McKeown, Clare Catherine (2019) *Optimisation and assessment techniques for clinical PET reconstruction.* PhD thesis

https://theses.gla.ac.uk/74344/

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses
https://theses.gla.ac.uk/
research-enlighten@glasgow.ac.uk

Optimisation and Assessment Techniques for Clinical PET Reconstruction

Clare Catherine McKeown
MEng, MSc

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy (PhD)

Department of Clinical Physics and Bioengineering
College of Medical, Veterinary and Life Sciences
University of Glasgow

August 2019

© Copyright 2019 by C. C. McKeown. All Rights Reserved.

Abstract

Introduction

Positron emission tomography (PET) is a molecular imaging technique; three-dimensional images of functional processes within the body are typically produced using iterative reconstruction methods. Image optimisation is a significant challenge, as the optimal combination of acquisition and reconstruction parameters is dependent upon image context and the clinical task, and there are thousands of possible combinations of selectable parameters. Furthermore, as PET technology continues to evolve, advances need to be optimised for different clinical indications. Manufacturers of PET imaging systems often suggest generic reconstruction strategies for tumour imaging of standard patients; however, imaging departments should validate and optimise reconstruction strategies for clinical applications of interest. This is particularly true when the use of such advances prove to be controversial, as remains the case for Point Spread Function (PSF) modelling.

Whilst many publications have assessed the effects of various reconstruction parameters, there is no established methodology for the assessment and optimisation of clinical reconstruction parameters. The primary aim of this thesis is therefore to develop a generic methodology to assess and optimise PET image reconstruction that can be applied to any clinical application.

The ability to detect small, low intensity lesions within the liver is critical for effective patient management; however, such lesions are challenging to identify in Fluorine-18 Fluorodeoxyglucose (^{18}F-FDG) PET imaging due to the relatively high liver background activity. Despite its clinical importance, there is no established optimal method for PET liver reconstruction. The secondary aim of this thesis is therefore to optimise image reconstruction for small liver lesion detection in ^{18}F-FDG-PET imaging for a specific PET system: the General Electric Medical Systems (GEMS) Discovery 690 PET-CT system.

Methods

Phantom studies were undertaken to assess the effects of varying acquisition and reconstruction parameters upon image noise, spatial resolution and lesion detection. The effects of slice overlap upon image quality were assessed to determine if the GEMS recommended setting was appropriate. The effects of Time of Flight (TOF), PSF, effective iterations, post-reconstruction filtering and voxel size on image quality were then assessed. A human observer study using patient data was also undertaken to determine if recommendations based on phantom data were applicable to clinical liver imaging. Different phantom acquisition and analysis techniques were compared and used to develop a generic

methodology that can be used to optimise PET acquisition and reconstruction for different clinical tasks. This methodology includes instructions for a patient observer study.

Phantom image acquisitions were largely designed to reflect liver imaging but can be adapted to other clinical scenarios. Phantom and patient work therefore led to the recommendation of a specific reconstruction strategy to optimise liver lesion detection.

Results

The use of a 23% slice overlap between image acquisition frames, as recommended by GEMS, was shown to produce acceptable image quality under routine clinical practice. An amendment to European guidelines was proposed to better reflect the relationship between image noise and slice overlap when calculating minimum patient injection activities.

A clinically relevant methodology for spatial resolution measurement was developed using activity concentrations and voxel values that reflect clinical imaging. Full width half maximum (FWHM) measurements were shown to be reliable and the use of a simplistic background activity correction method was assessed and justified.

The effects of reconstruction parameters upon image noise and clinical spatial resolution were assessed using phantom data. PSF was shown to degrade spatial resolution at low iterations (<180) when applied without TOF. Furthermore, noise analyses revealed the GEMS Gaussian filter implementation did not function as intended at certain filter widths; larger voxels demonstrated greater noise levels than smaller voxels. This was unknown to GEMS engineers and was a novel finding.

Qualitative phantom assessments concluded applying TOF and PSF together, with no filtering and approximately 108 iterations, was optimal for small lesion detection. Joint analysis of Hot Contrast Recovery Coefficient (HCRC) and noise were shown to be a better predictor of observer lesion detection preferences than signal-to-noise ratios (SNR). Furthermore, a novel SNR calculation (based on region-to-region noise) was shown to be a better predictor of human observer preference than traditional SNR calculations (based on voxel-to-voxel noise).

Reassuringly, reconstruction parameters suggested by GEMS (54 iterations, 4mm filter, 3.65mm voxels) were found to be close to optimal. The patient observer study (n = 30) did not demonstrate any statistically significant differences in lesion detection between the GEMS suggested reconstruction and three progressively sharper reconstructions based on the phantom results. However, results did suggest the use of a sharper reconstruction (54 iterations, no filtering, 2.73mm voxels) may improve lesion detection.

Conclusions

A four-step generic methodology for optimising PET acquisition and reconstruction is proposed by this thesis. This methodology includes instructions for selecting an appropriate slice overlap for image acquisition. Phantom acquisition and analysis techniques are included for assessing spatial resolution, image noise and lesion detection in a clinically relevant manner. Recommendations are also made for conducting a patient observer study.

This thesis further concludes that all ^{18}F-FDG oncology patient images acquired using the GEMS Discovery 690 PET-CT system should be reported using two reconstructions in tandem. The GEMS suggested reconstruction should be used for general image interpretation. The sharper reconstruction should be used for liver lesion detection.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Firstly, I would like to thank my supervisors, Dr Gerry Gillen and Dr Mary Frances Dempsey, for their constant support, guidance and motivation over the course of my PhD studies. I will be forever grateful to you both for sacrificing so much of your personal time for this project. In particular, I would like to thank Gerry for continuing in his supervisory role after taking early retirement.

I also owe a debt of gratitude to Dr Caroline Findlay for all of her support and advice, particularly during the write-up phase. Thanks also to Dr John Shand and Dr David Colville for giving up their time to assist with my research.

I would like to extend my thanks to all my colleagues in both the PET and Nuclear Medicine departments in Gartnavel General Hospital for all their support over the years. In particular, thanks to Dr Colin Brown for his part-time PhD camaraderie, and for proving it was actually possible to reach the finishing line!

I must also thank my wonderful friends outside of work for helping me keep my sanity over the last few years. Thank you for your understanding, your words of encouragement…and even the delicious food parcels that helped me survive the last few months of writing!

To my mum, Margaret, and my sister, Frankie – I am very lucky to have you both in my corner. Thank you for all your patience and understanding over the last few years, and for always knowing when chocolate and cider were required. I don't think I could have finished this thesis without your support. I promise to repay you both, somehow!

Finally, I would like to thank my dad, Billy… I wouldn't have made it this far if you hadn't helped me with my physics homework all those years ago. I will raise a glass to you at my graduation celebrations!

Author's Declaration

I declare that this thesis represents my own work except where referenced to others.

I declare that this thesis does not include work forming part of a thesis presented successfully for another degree.

Clare McKeown

August 2019

Definitions and Abbreviations

^{18}F	Fluorine-18
^{22}Na	Sodium-22
^{68}Ga	Gallium-68
^{68}Ge	Germanium-68
2D	Two Dimensional
3D	Three Dimensional
AC	Attenuation Correction
AC	Activity Concentration
AAC	Absolute Activity Calibration
ACRIN	American College Radiology Network
AFROC	Alternative Free Response Receiver Operator Characteristic
APD	Avalanche Photodiode
ARSAC	Administration of Radioactive Substances Advisory Committee
AUC	Area Under Curve
BGO	Bismuth Germanate
BMI	Body Mass Index
Bq/ml	Becquerels per millilitre
BV	Background Variation
CCRC	Cold Contrast Recovery Coefficient
COV	Coefficient of Variation
cps	Counts per second
CQIE	Centre of Quantitative Imaging Excellence
CRC	Colorectal Cancer
CRC	Contrast Recovery Coefficients
CT	Computed Tomography
CTAC	Computed Tomography Attenuation Correction
CTW	Coincidence Timing Window
D690	General Electric Discovery 690 PET-CT System
DLBCL	Diffuse Large B Cell Lymphoma
DOI	Depth of Interaction
EANM	European Association of Nuclear Medicine
EARL	European Association of Nuclear Medicine Research Limited
FBP	Filtered Back Projection
FDG	Fluorodeoxyglucose
FN	False Negative

FOV	Field of View
FP	False Positive
FPF	False Positive Fraction
FROC	Free Response Receiver Operator Characteristic
FWHM	Full Width Half Maximum
GCAT	Gamma Camera Assessment Team
GEMS	General Electric Medical Systems
HCC	Hepatocellular Carcinoma
HCRC	Hot Contrast Recovery Coefficient
HD	High Definition
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronic Engineers
IR	Image Roughness
JAFROC	Jackknife Alternative Free Response Receiver Operator Characteristic
LAC	Linear Attenuation Coefficient
LLF	Lesion Location Fraction
LOR	Line of Response
LROC	Localisation Receiver Operator Characteristic
LSO	Lutetium Orthosilicate
LYSO	Lutetium Orthosilicate doped with Yttrium
MBq	Megabecquerels
MDA	Medical Devices Agency
MLEM	Maximum Likelihood Expectation Maximisation
MRI	Magnetic Resonance Imaging
MTV	Metabolic Tumour Volume
NAC	Non Attenuation Correction
NCRI	National Cancer Research Institute
NECR	Noise Equivalent Count Rate
NEMA	National Electrical Manufacturers Association
NIST	National Institute of Standards and Technology
NLF	Non-Lesion Location Fraction
NPV	Negative Prediction Value
OSEM	Ordered Subsets Expectation Maximisation
PERCIST	Positron Emission tomography Response Criteria in Solid Tumours
PET	Positron Emission Tomography
PHA	Pulse Height Analysis
PMT	Photomultiplier Tube

PPV	Positive Prediction Value
PSF	Point Spread Function
PV	Partial Volume
QA	Quality Assurance
QIBA	Quantitative Imaging Biomarkers Alliance
ROC	Receiver Operator Characteristic
ROI	Region of Interest
SD	Standard Deviation
SharpIR	Sharp Iterative Reconstruction
SiPM	Silicon Photomultiplier
SNMMI	Society of Nuclear Medicine and Molecular Imaging
SNR	Signal to Noise Ratio
SUV	Standardised Uptake Value
TLG	Total Lesion Glycolysis
TN	True Negative
TOF	Time of Flight
TP	True Positive
TPF	True Positive Fraction
VPFX	Vue Point with Time of Flight
VPHD	Vue Point High Definition
wAFROC	Weighted Alternative Free Response Receiver Operator Characteristic
WCC	Well Counter Correction

Chapter 1 : Introduction

1.1 Cancer

Cancer is a disease characterised by mutations that disrupt the control mechanisms for cell division, growth and death. These abnormal cells typically have the ability to sustain uncontrolled proliferation and can form malignant tumours [1]. As the disease progresses, primary tumours can invade surrounding tissues and metastasise to other sites in the body, causing secondary tumours. Malignant tumours can grow at the expense of healthy cells by depriving them of essential nutrients, and vital organs can be prevented from functioning properly. Eventually, this may lead to death [2]. In 2012, there were approximately 8.2 million deaths from cancer worldwide (representing one in every six deaths) with 14.1 million new cancer diagnoses [3]. By 2018, these numbers had risen to approximately 9.6 million deaths (17% increase in six years) and 18.1 million new diagnoses (28% increase) [4]. Global cancer incidence is expected to continue to increase due to population growth and aging [5], and will soon become the most common cause of death worldwide [6].

Cancer is more likely to respond to treatment when it is detected early, resulting in increased probability of survival and reduced treatment costs [7]. The ability to detect small lesions during the early stages of disease is therefore critically important [8]. Imaging has a key role in lesion detection and is widely used for assessing prognosis and determining the most appropriate course of treatment [9]. Imaging also has significant roles in determining whether to continue, change or abandon treatment for individual patients, and in clinical trials assessing the efficacy of new treatment strategies [6].

1.2 Positron Emission Tomography in Cancer Management

Positron emission tomography (PET) is a nuclear, molecular imaging technique that produces three-dimensional images of functional processes within the body. By identifying changes in the body at a molecular level, PET imaging may detect the early onset of disease before it is evident on structural imaging modalities, such as Computed Tomography (CT) or Magnetic Resonance Imaging (MRI) [10]. The use of short-lived positron-emitting radiopharmaceuticals allows biologically important molecules to be labelled without significantly changing their behaviour [11]. These radiolabelled 'tracers' are typically introduced into the patient's body by intravenous injection.

The ability to measure cellular glucose metabolism is important in oncology because many types of cancer cells have an abnormally high dependence on glucose, known as the 'Warburg effect' [12], [13]. Fluorine-18 coupled to a glucose analogue (^{18}F-Fluorodeoxyglucose, commonly abbreviated to ^{18}F-FDG) is therefore one of the most commonly used PET tracers. ^{18}F-FDG follows a similar metabolic pathway to glucose, except it is retained in the cells in proportion to the rate of glycolysis [10]. This tracer accumulation allows localisation of these cells in the PET images. ^{18}F-FDG PET has been shown to be a sensitive imaging modality for a variety of cancers [14].

The development of hybrid PET-CT clinical systems in the early 2000's, combined with the availability of commercial ^{18}F-FDG supplies, led to the rapid establishment of PET-CT in clinical cancer management [15]. FDG-PET is now routinely used in the diagnosis, staging, response evaluation, restaging, follow-up and radiotherapy treatment planning of many types of cancers [16]. However, PET imaging has its limitations; e.g. PET images have poorer spatial resolution than anatomical imaging techniques such as MRI and CT, which can make the detection of small lesions challenging. Consequently, PET technology has evolved rapidly in recent years as new hardware and software solutions have been developed.

1.3 Basic Principles of Positron Emission Tomography

The basic principles of PET data acquisition and storage, as illustrated in Figure 1.1, are well known in the field and are therefore only briefly summarised in this thesis.

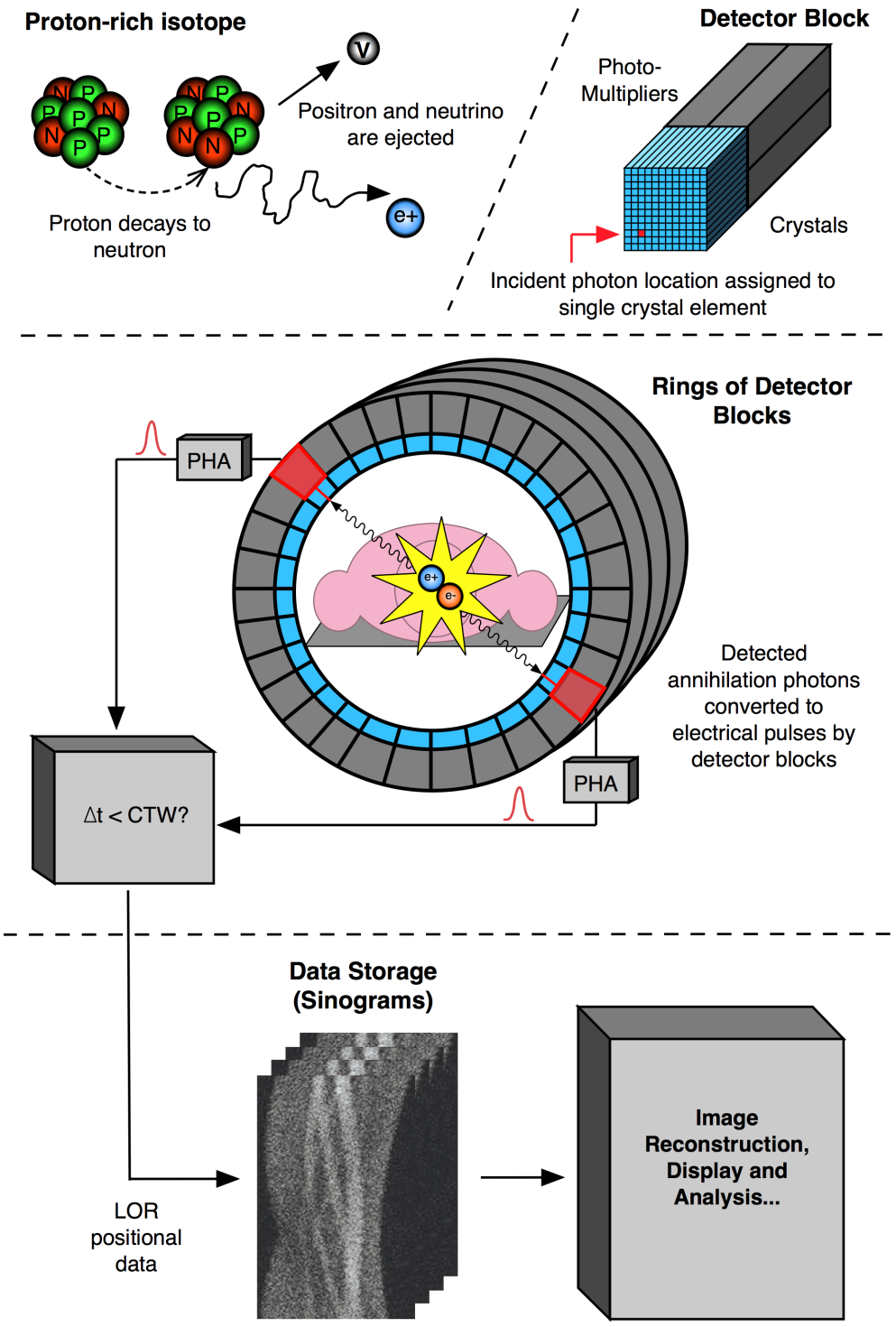


Figure 1.1: Basic principles of PET data acquisition and storage

1.3.1 Positron Emission and Annihilation

An unstable nucleus in a proton-rich radionuclide converts a proton into a neutron as it decays to a stable state and releases surplus positive charge by ejecting a positron (a neutrino is also ejected but is not of use in PET imaging). The positron travels a short distance through tissue, losing kinetic energy as it collides with surrounding atoms. Once it has lost almost all of its kinetic energy, the positron undergoes mutual annihilation with an

electron and their rest masses are converted into a pair of annihilation photons. To conserve energy, each photon has energy equal to the rest mass of the positron/electron, which is 511keV. To conserve momentum, the photons are emitted at almost 180° to each other. Detection of these anti-parallel 511keV annihilation photons is used to construct the PET images.

1.3.2 Detection System

PET detectors use scintillation crystals such as bismuth germanate (BGO) or lutetium Orthosilicate (LSO) to convert incident annihilation photons into light photons. Photomultiplier tubes (PMTs) are typically used to convert these light photons into electrical pulses, with heights proportional to the energy of the incident annihilation photons. However, the advent of integrated PET and MRI systems has seen the development of alternatives to PMTs (as they are susceptible to magnetic fields and therefore incompatible with MRI). Such alternatives include avalanche photo detectors (APDs) and silicon photomultipliers (SiPMs).

PET gantries generally comprise rings of detector blocks, which surround the patient. An example detector block is shown in Figure 1.1 and consists of a crystal array coupled to four PMTs. The magnitude of the output pulses from each of the four PMTs are used to calculate which crystal in the array detected the annihilation photon. Clinical whole-body PET scanners have ring diameters (transaxial Field of View, FOV) of approximately 70 – 80cm. Multiple detector rings are used to create gantries with axial FOVs of approximately 15cm to 25cm.

Whole-body PET scans are typically acquired with overlapping acquisition frames (or beds) to account for reduced sensitivity at axial FOV edges. Data in overlapped regions can then be combined to improve the statistics. This improves quantitative accuracy but increases the required scan time [17]. Sensitivity and overlapping frames are discussed in more detail in Chapter 4, which also includes an investigation into how overlap size affects image quality.

1.3.3 Coincidence Detection

Electrical pulses from the detection system are fed into Pulse Height Analysis (PHA) circuitry. If the PHA determines the detected photon’s energy is within the energy acceptance window for 511keV, a timing pulse is generated; otherwise, the detected event is rejected. The energy acceptance window is largely dictated by the energy resolution of the crystal: the better the energy resolution, the narrower the accepted range of photon energies.

Two detected photons are considered to be coincident (i.e. from the same annihilation event) if their timing pulses are both generated within a Coincidence-Timing Window (CTW). The size of this timing window is dependent on the decay time of the crystal. BGO crystals typically have CTWs of 10-12ns, while faster LSO crystals typically have CTWs of 4-6ns [18]–[20].

The location of the detected annihilation event is known to lie somewhere along the line between the two detection crystals, known as a Line of Response (LOR) (sometimes referred to as a Tube of Response [21] or a Volume of Response [28]).

1.3.4 Sinograms

Raw PET data are stored in two-dimensional arrays known as sinograms. Each sinogram bin represents a LOR between two crystals and the value of each bin represents the number of events detected along the LOR. Each row in a sinogram represents the data collected from parallel LORs at a given radial offset angle. A simple sinogram example is shown in Figure 1.2: four LORs with different offset angles originate from an off-centre point source and are stored in the corresponding sinogram bins. When more radial angles contribute to the sinogram, the offset point source produces a sinusoidal pattern (shown in grey).

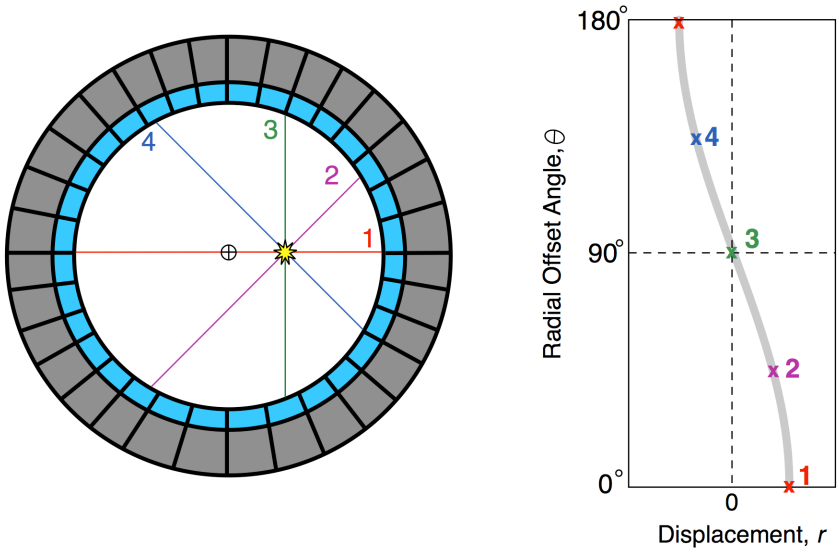


Figure 1.2: Simple sinogram explanation

LORs are typically defined by two coordinates: the transaxial angle between the LOR and the x-axis (θ) and the radial distance of the LOR from the centre axis of the detector ring (r). In 3D PET, LORs are also defined by coordinates that account for the axial angle between the LOR and the z-axis (ϕ) and the distance of the LOR from the centre of the z-axis (z) [23].

1.4 PET Image Reconstruction

The aim of PET image reconstruction is to produce quantitatively accurate cross-sectional images of the radiotracer's distribution within the FOV using the detected LORs. This is challenging because the acquired PET data is inherently noisy:

- The amount of radioactivity that may be injected into the patient is subject to restrictions which limit the risks of radiation exposure: for example, the maximum activity of ^{18}F-FDG that may be injected for whole-body oncology exams in the UK is 400MBq [24].
- The length of acquisition time is limited: current PET systems can only image 15-25cm sections of the patient at a time (multiple bed positions are required for whole-body scans), and the possibility of patient motion increases with acquisition time (which may cause motion artefacts).

These restrictions limit the number of annihilation events available for detection, leading to increased statistical noise. Furthermore, the sensitivity of PET systems is inherently poor due to the limited axial detector coverage and isotropic nature of the emissions; less than 1% of coincidence effects are detected [25].

Image reconstruction in PET is an *'inverse problem'*: measured LORs are used to calculate the radioactivity distribution from which the LORs arose. Image reconstruction in PET is also an *'ill-posed problem'*. A problem is ill-posed if at least one of the following conditions are met [26]:

1. Solution does not exist (not an issue in practical applications), or
2. There is no unique solution, or
3. The solution is unstable: small errors in the measurement data may lead to large errors in the solutions (also known as an *'ill-conditioned problem')*.

The PET image reconstruction problem meets conditions 2 and 3. The uniqueness of the solution may fail due to the low-count, noisy nature of clinical PET data. Small changes in the measured data can be propagated through reconstruction operations to become larger changes in the final image [27].

There are two basic approaches to image reconstruction: analytical methods and iterative methods [28]. Analytical methods typically assume there is a unique solution to the measured projections; however, the presence of noise in the measured data means that there are normally a number of possible solutions. Iterative reconstruction techniques determine the 'most likely' solution by using a feedback loop.

1.4.1 Filtered Back Projection

Analytic methods use the mathematics originally developed for Computed Tomography (CT) that relate measured projections to the object's activity distribution. They offer a direct mathematical solution for the formation of an image [22]. The most common analytical technique is filtered back projection (FBP).

Simple backprojection is shown in Figure 1.3 (A). A set of projections are measured from an object in the FOV (each projection angle forms a line in the sinogram). Each measured projection is *back projected* along the image matrix. Counts are spread along the path from which they were originally acquired, resulting in a blurry version of the ideal image. Such artefacts are known as 'star' artefacts [29], [30], and can be reduced by applying a ramp filter to each projection prior to back projection. Ramp filters amplify high frequency noise and are therefore typically combined with low-pass smoothing filters. This *filtered back projection* reduces the noise in the final image at the expense of spatial resolution (Figure 1.3 (B)) [30]. Different degrees of filtering lead to different noise and resolution image characteristics.

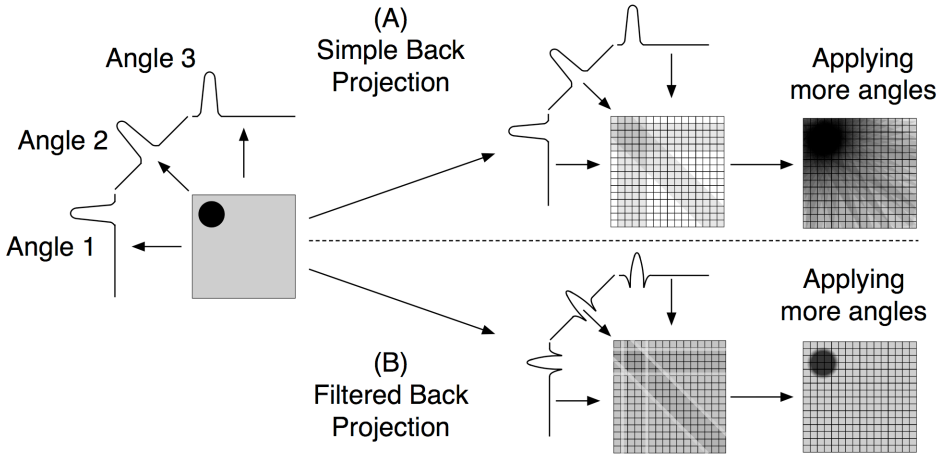


Figure 1.3: Simple example of filtered back projection

1.4.2 Iterative Reconstruction

Iterative reconstruction methods offer improvements over analytical methods because they can account for the noise structure in the measured projections and can use a more realistic model of the imaging system [22]. Such methods model the data collection process in the PET scanner and attempt to find the image that is most consistent with the measured projection data [31]. Iterative methods are computationally more intensive than analytical methods; however, advances in computation speed and the availability of faster algorithms has enabled these techniques to be used clinically.

To determine the 'best' solution, criteria are defined to measure the goodness of fit between an estimated image and the measured data. A feedback loop allows sequential adjustments to be made to the estimated image to improve the fit with the measured data [32]. The most common iterative reconstruction method used in today's clinical PET systems is the Ordered Subsets Expectation Maximisation (OSEM) algorithm [17], which is based upon the Maximum Likelihood Expectation Maximisation (MLEM) algorithm.

1.4.2.1 Maximum Likelihood Expectation Maximisation

MLEM is a statistical reconstruction method that uses a model of the imaging system, known as a *system matrix*, to define the relationship between image space and sinogram space [22], [33]. The system matrix describes the set of probabilities, based on Poisson statistics, that a pair of annihilation photons emitted from image voxel *j* will be detected in the sinogram bin *i*, for all (*i, j*) pairs [34]. This conditional probability is referred to as *likelihood*, which is a general statistical measure that is maximised when the difference between the measured and estimated sinograms is minimised [28], [35]. The system matrix should account for physical factors which affect the probability of an annihilation pair being detected in a particular sinogram bin; for example, the uneven spacing of the projections due to curved detector geometry and the crystal block design [36]. Other factors considered by the system matrix, such as positron range, photon non-colinearity and data corrections are discussed later in this chapter. The quality of reconstructed images is critically dependent upon the accuracy of the system matrix [37].

Figure 1.4 outlines the basic principles of MLEM reconstruction. An initial estimated image of radioactivity distribution is provided to begin the iteration sequence. This may be a blank image, a uniform image or a simple FBP reconstruction [28], [38]. The algorithm *forward projects* this estimated image to produce a set of estimated projections. Forward projection is the inverse of back projection: all voxels that are intersected by a particular LOR are summed to fill the corresponding sinogram element. These estimated projections represent the data that would have been acquired had the estimated image been a true representation of the radioactivity distribution. The estimated projections are then compared with the measured projections and the initial estimated image is adjusted based on these differences. The new estimated image is then forward projected, and the full process is reiterated. Each successive iteration should produce an image which represents the true activity distribution more closely than the previous iteration: i.e. the estimated image should *converge* towards the true image [35]. However, the ill-posed nature of the reconstruction problem, combined with limitations in the system matrix's accuracy, means a fully converged image is likely to have excessively high noise [39]. Such algorithms are typically stopped after a specified number of iterations to achieve an acceptable trade-off between

accuracy and noise. Post-reconstruction filtering may also be employed to reduce the noise, as discussed later in this chapter.

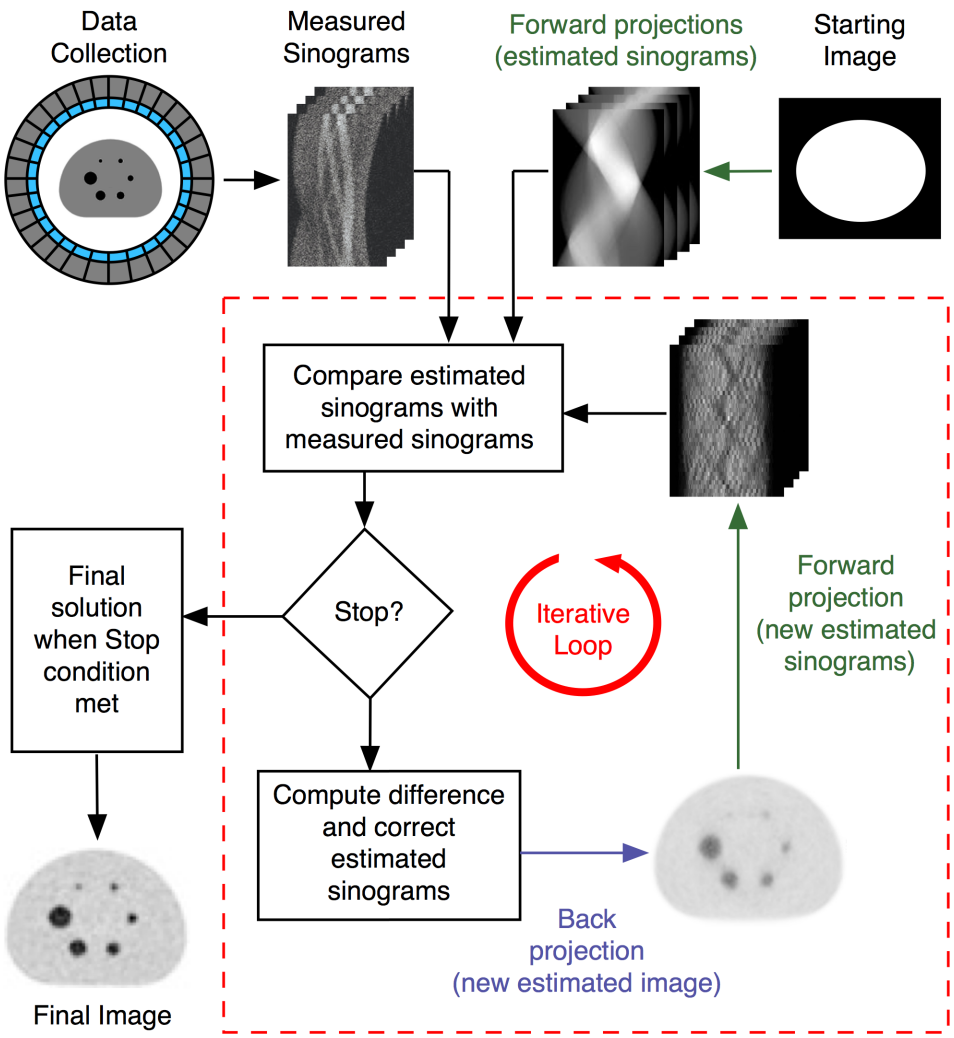


Figure 1.4: Principles of MLEM algorithm

A major disadvantage of MLEM reconstruction is the computation time required, as each iteration requires a forward and backward projection operation, and many iterations are required to reach convergence. Fifty MLEM iterations takes approximately 100 times as long as a FBP reconstruction [32].

1.4.2.2 Ordered Subsets Expectation Maximisation

OSEM is a variant of MLEM that accelerates the reconstruction process by grouping the projection data into blocks, called subsets, within each iteration. Forward projection and back projection are performed on each subset in turn. The resultant reconstruction for one subset is used as the starting estimate for the next subset. One full OSEM iteration is complete once all subsets have been through the forward/back projection step. Subsequent OSEM iterations can then be performed using the end result of the previous OSEM iteration.

This approach was shown to accelerate convergence by a factor proportional to the number of subsets [40]. *Effective iterations* in OSEM are the product of subsets and iterations [41]: for example, three OSEM iterations using 18 subsets is virtually equivalent to performing 54 MLEM iterations, whilst taking the same amount of time as only three MLEM iterations. OSEM is typically applied with small numbers of iterations (e.g. less than five) with a higher number of subsets (e.g. more than ten).

The number of effective OSEM iterations applied affects the degree of convergence and the noise level in the image: an increased number of iterations achieves better convergence and contrast recovery, but at the cost of a higher level of noise. Convergence rates are also affected by the image distribution, and smaller objects require more iterations to reach convergence than larger objects [42]. The number of subsets/iterations used should therefore be tailored to the desired convergence/noise level and the specific diagnostic task [43].

1.4.3 Matrices and Voxel Sizes

PET image reconstruction can be performed with different matrix sizes, which determine the size of the voxels that make up the reconstructed image. The frame length and the number of transverse image slices in the frame, both of which are typically fixed, determine the size of the voxel in the z-direction. Smaller voxel sizes can be achieved by increasing the matrix size in the transaxial plane and/or reducing the transaxial FOV. This may be desirable, for example, in head and neck imaging, where both the area of interest and the potential lesions may be small. Smaller voxel sizes can improve PET spatial resolution; however, as voxel size decreases, the number of reconstructed events that contribute to each voxel also decreases, and hence the statistical noise in each voxel increases [17]. Therefore, a trade-off exists between spatial resolution and image noise when considering voxel sizes.

1.4.4 Post Reconstruction Filtering

Post reconstruction image filtering, or smoothing, is used to reduce the effects of noise on image interpretation and analysis [44]. Smoothing operations average local voxel values to reduce the effect of voxel-to-voxel variations in the reconstructed image [45]. This has the same effect as a low-pass filter, which removes high frequencies from the data. Smoothing reduces noise but this is at the expense of spatial resolution; high frequency detail is lost to the ‘blurring’.

Post-filtering reconstructed images with a Gaussian kernel is the most common filtering method in PET clinical practice [17], [46]. This type of filter produces a weighted-average of

each voxel's neighbourhood, with the average value weighted towards the central voxel values. Post reconstruction Gaussian filters are described in more detail in Chapter 6.

1.4.5 Data Corrections

The following corrections are typically performed within the iterative reconstruction loop in order to improve overall accuracy, uniformity of convergence [36] and preserve Poisson statistics [47]–[49].

1.4.5.1 Scatter Correction

511keV photons are likely to undergo Compton scattering, which can change their direction without significantly reducing their energy [50]. The limited energy resolution of PET crystals means that some scattered photons will be accepted. Incorrect LORs can therefore be assigned, and true points of annihilation are lost. The number of scattered events registered depends on the size of the object being imaged (larger objects cause more photons to be scattered) and the width of the energy acceptance window. The proportion of accepted scatter coincidences is known as the scatter fraction, and can be as high as 50% in 3D-mode acquisition [51].

Scatter correction techniques include empirical measurement, subtraction using multiple energy windows, convolution, and scatter distribution modelling during iterative reconstruction techniques [29], [50]. Model-based scatter correction techniques, such as Monte Carlo simulations and Klein-Nishina formula-based corrections, are computationally intensive but are widely used in modern PET systems [30].

1.4.5.2 Randoms Correction

A random coincidence may be registered as a result of single photons from two separate annihilation events arriving at detectors within the CTW. The true LORs are missed, and an incorrect LOR is registered instead. The rate of randoms increases with the square of the activity and is directly proportional to the size of the CTW.

Random event rates can be estimated by using a second coincidence window, which delays the input from one of the detectors by a time period that is greater than the true coincidence window. Random correction is achieved by subtracting the counts measured in the delayed window from those measured in the true window [52]; however, this increases statistical noise in the subtracted data [53]–[55]. Another correction method is singles-based randoms correction, where the number of random events is estimated using the number of single events detected. This method has been shown to yield substantial improvements over the delayed-window method [53]–[55].

1.4.5.3 Attenuation Correction

Photon attenuation is the most significant cause of PET image quality degradation and affects both visual interpretation and quantitative accuracy. The degree of attenuation experienced by annihilation photons depends on the depth of the annihilation event within the attenuating structure and the linear attenuation coefficient (LAC) of the surrounding tissue. Activity at the periphery of an object is subjected to less attenuation than activity at depth. This can artificially reduce the detected counts for deep lesions.

Attenuation of a particular LOR is independent of the position of the annihilation event along the LOR. Attenuation can therefore be corrected by creating an attenuation map, based on LACs along each LOR, and using it to correct the detected counts. Dual modality PET-CT scanners perform CT imaging of the patient to obtain an attenuation map. This also provides reporting clinicians with high-resolution structural data that can be used in conjunction with the functional PET data for anatomical localisation. CT data must be processed prior to attenuation correction, as the CT photons (typically 70keV→120keV) experience greater attenuation than 511keV photons. CT Hounsfield units are translated into 511keV LACs, typically using a combination of segmentation and scaling [56]–[58]. Spatial resolution of the CT attenuation correction (CTAC) map is then smoothed to a resolution that matches the PET data [59]. Attenuation correction factors are generated by forward projecting along the LORs through the CTAC map and applied to acquired data during the reconstruction process.

1.4.5.4 Normalisation Correction

Thousands of scintillation crystals are used in PET detection systems. Non-uniformities in individual crystals, geometric variations and detector electronics contribute to variations in the efficiency of the system LORs [28], [59]. Normalisation is performed by exposing every possible LOR to a uniform positron-emitting source, for example by rotating a positron-emitting rod source around the detector ring or by placing a uniform cylindrical phantom at the centre of the FOV [60]. Correction factors, or normalisation coefficients, are calculated for each individual LOR and applied to acquired PET data. Normalisation data is often incorporated into the system matrix [36].

1.4.5.5 Dead Time Correction

The ‘dead time’ of a system describes total time taken to detect a photon, generate an electronic pulse, determine the energy and spatial position of the photon, record a count and determine that a coincidence event has occurred [30], [61]. During this time, the detection system is unable to process subsequent events. The detection system therefore has a limit to the rate at which events may be processed [29]. ‘Dead time losses’ describes

the count losses experienced due to scintillation decay time and the limited speed of processing electronics. The effects of these limitations intensify as the count-rate increases. Mathematical corrections can be used to compensate for dead time effects by applying multiplicative factors to the measured counts [30]. However, using crystals with faster decay times and using faster electronics can reduce the system's dead time directly. Dead time data is often incorporated into the system matrix [36].

1.4.5.6 Decay Correction

A whole-body scan may take over 30 minutes to acquire. The radionuclides used in PET have short half-lives and therefore counts acquired in each bed position must be corrected for radioactive decay.

1.5 Quantitative PET

A major advantage of PET imaging is the ability to quantify the accumulated radiopharmaceutical within tissue, provided that the corrections described in the previous section are performed correctly, allowing lesions to be classified according to their metabolic rate [28]. Although visual inspection of PET images is very important for diagnosis and response assessment, many studies have shown that quantification allows objective assessment of lesion characterisation, prognosis and response to treatment [62].

1.5.1 Well Counter Calibration

The PET imaging system must be calibrated to convert measured count densities into activity concentrations. Well Counter Calibration (WCC) matches the radioactivity measured by the radionuclide calibrator (used to measure patient injections) to the radioactivity measured by the PET scanner [63], [64]. WCC correction factors are used to convert image voxel values into a measure of absolute activity concentration per voxel (e.g. Becquerels per millilitre, Bq/ml) [65]. Accurate WCC is crucial to ensure the accuracy of the PET data. WCC is discussed in more detail in Chapter 5, which also examines the effects of reconstruction parameters on voxel values.

1.5.2 Standardised Uptake Values

Absolute quantification is complex and therefore is difficult to achieve in routine clininal imaging. However, a semi-quantitative measurement known as the Standardised Uptake Value (SUV) is widely used:

$$SUV = \frac{Measured\ Activity\ Concentration\ in\ Tissue}{Injected\ Activity / Body\ Size} \qquad \text{Equation 1.1}$$

SUVs are measured by drawing a region around the tissue of interest (e.g. tumour) using image analysis software. The measured radioactivity within the region is then normalised to the average radioactivity concentration throughout the patient's body. Measurements can be based on body weight, lean body mass or body surface area, with body weight the most frequently used [66].

There are a number of different parameters used for expressing SUVs:

- SUV_{max}: based on the maximum voxel value within the defined region
- SUV_{mean}: based on the average voxel value within the defined region
- SUV_{peak}: based on the local average of a small, fixed area of voxels which surround the voxel with the highest activity [6]

SUV_{max} is currently the most commonly used SUV measurement because it is less observer dependent and more reproducible than the other measurements [66], [67]. However, the 'single voxel' nature of SUV_{max} makes it vulnerable to statistical noise in the image data [68].

The following factors can introduce variability into SUV measurements in ^{18}F-FDG studies:

- Region of interest (ROI) placement, shape and size (i.e. selected by visual judgement)
- Partial volume effects (discussed later in this chapter)
- Image reconstruction parameters
- Uptake period between radiopharmaceutical administration and image acquisition
- Blood glucose level
- Patient size

SUVs should therefore be used with caution when assessing malignancy [14], [59].

1.6 PET Spatial Resolution Limitations

Spatial resolution is a measure of an imaging system's ability to accurately distinguish between two close together objects and observe their details [69]. The point spread function (PSF) is used to describe the response of the imaging system to a point source. Spatial resolution is typically described by the full width half maximum (FWHM) of the PSF at various positions within the FOV [70]. Figure 1.5 shows an ideal point source and its frequency spectrum (green lines): all spatial frequencies are required for an accurate point source representation [71]. A detection system with a typical Gaussian PSF would produce the PSF shown by the red lines in Figure 1.5. Its frequency spectrum shows the loss of high frequency components, which results in the loss of fine detail in the reconstructed images.

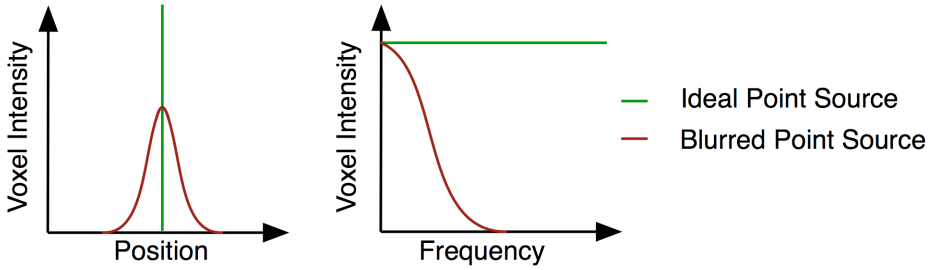


Figure 1.5: Point source in spatial and frequency domains

There are a number of factors that affect the spatial resolution of PET images.

1.6.1 Emission Process

The distance travelled by a positron prior to annihilation introduces an inherent spatial resolution limitation. Coincidence detection relates to the site of annihilation rather than the site of the positron's parent nucleus. Higher energy radionuclides have greater positron ranges in tissue than lower energy radionuclides. For example, ^{18}F (0.6mm mean positron range) will produce images with better spatial resolution than ^{15}O (1.5mm mean positron range).

A second inherent spatial resolution limitation results from non-colinearity of the annihilation photons. Positrons typically have some residual momentum at the point of annihilation. This causes a small deviation from 180° between the anti-parallel annihilation photons, which in turn causes the LOR between the two detectors to be slightly displaced from the point of annihilation. The maximum deviation from 180° is ±0.25° [72]. The effect of non-colinearity worsens as the size of the detector ring increases. For an 80cm FOV, non-colinearity amounts to ≈ 1.8mm [61].

1.6.2 Detection Process

The size of the scintillation detector elements has a major impact on spatial resolution: smaller detectors create thinner, more precise and more numerous LORs, improving sampling of the object in the imaging FOV. Smaller detectors can therefore increase spatial resolution [73]; however, this must be optimised with respect to sensitivity [74]. Furthermore, detector response to a point source is dependent on the source's position within the FOV, as shown in Figure 1.6 (A). The response to a point source mid-way between two detectors is triangular, with a FWHM equal to half of the detector width. The response worsens as the point source moves towards either of the two detectors (i.e. closer to the FOV edge) and becomes trapezoidal in shape, with a FWHM matching the detector width [29], [75]. Therefore, for a detector width d, resolution is ≈ $d/2$ at the centre of the FOV and ≈ d at the face of the detectors.

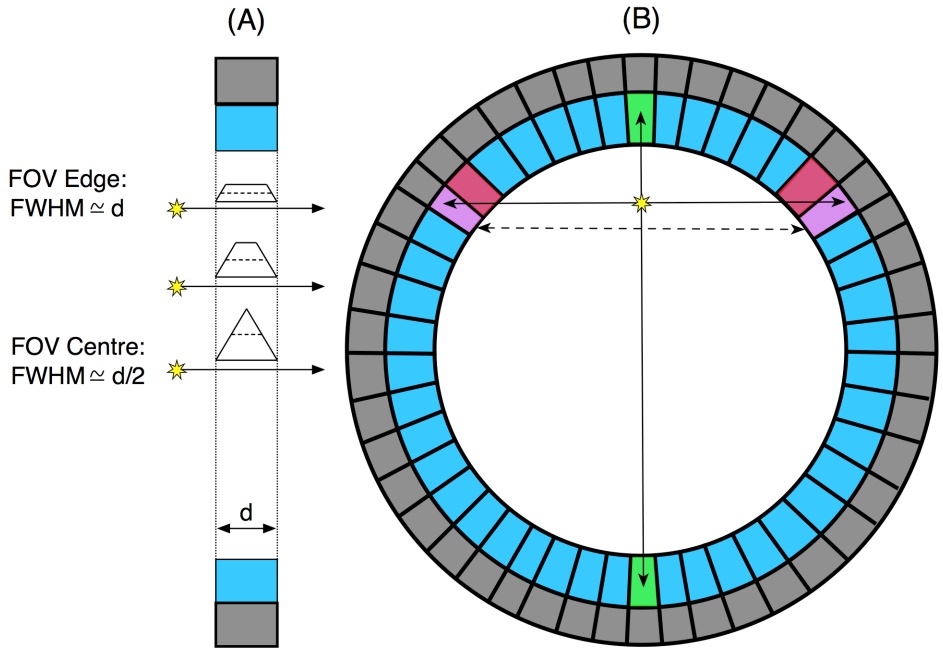


Figure 1.6: Detector effects on spatial resolution

Parallax effects, or radial elongation, also contribute to spatial resolution variability throughout the FOV, as shown in Figure 1.6 (B). Photons incident perpendicularly on a particular crystal are more likely to be absorbed within that crystal (shown in green), producing a true LOR. However, photons that enter a particular crystal at an acute angle are more likely to penetrate into the neighbouring crystal(s) before being absorbed (shown in pink). The depth-of-interaction (DOI) within the crystals is unknown and is not accounted for, resulting in an incorrect LOR (dotted line). The FWHM of point sources located nearer the FOV edges are therefore broadened.

1.6.3 Partial Volume Effects

Partial Volume (PV) effects refer to phenomena that cause voxel intensities to differ from what they should be: smaller objects can appear to have lower activity concentrations than larger objects with equal activity concentrations [28]. PV effects lead to underestimation of the activity concentration of lesions smaller than twice the system spatial resolution FWHM [17], [32]. The extent of PV effects depend on the lesion size, the contrast between the lesion and its background, and the system spatial resolution [76]. The term ‘partial volume effect’ typically refers to two distinct phenomena [77]:

1. Three-dimensional image blurring introduced by the finite spatial resolution of the imaging system
2. Sampling effect of voxel sizes (tissue fraction effect)

Image blurring as a result of limited spatial resolution causes ‘spillover’ between regions, which in turn causes small sources to appear larger and less intense. As spatial resolution degrades towards the edges of the FOV, PV effects also become more significant at FOV edges [17].

Sampling the radiotracer distribution onto a voxel grid also causes PV effects. Most voxels will contain different types of tissues (tissue fraction effect [77]). Each voxel’s signal intensity is the average of all underlying tissues within the voxel. Voxels around the edge of a source will contain both source and background tissue. Large voxels are more likely to contain a mixture of different tissue types than small voxels. Averaging source and background tissue signals causes the source to appear larger and less intense than it should be, which contributes to the ‘spilling out’ effect. Activity from background tissue can also ‘spill in’ to the source tissue, which may partially compensate for the ‘spill out’ effect, depending on the background activity concentration.

The use of smaller voxels reduces the tissue fraction effect, and therefore minimises PV effects; however, this must be balanced with the associated increase in statistical noise. Motion correction techniques, such as respiratory and cardiac gating, can be used during acquisition to minimise blurring caused by physiological movement. There are many methods for PV correction currently under investigation, but these are yet to gain widespread acceptance in clinical PET imaging. PV correction can be applied as part of the reconstruction algorithm: for example, the detector PSF response can be modelled and incorporated into the reconstruction algorithm. This type of correction is described in more detail later in this chapter. Alternatively, PV correction can be applied as a post-reconstruction technique: for example, high-resolution structural images (such as CT images) can be used to transfer high frequency information [78].

1.7 Developments of New PET-CT Technology

There have been many technological advances made in commercially available, clinical PET-CT systems in recent years. Some of these advances are beyond the scope of this thesis; for example, the use of solid state photodetectors instead of PMTs [79], [80] and regularised reconstruction algorithms that incorporate penalty terms to suppress noise while reaching convergence [81]. However, there are two major developments that are of particular interest to this thesis: Time of Flight imaging and Point Spread Function modelling.

1.7.1 Time of Flight

In conventional PET imaging, a LOR contains no positional information for location of the annihilation event along the line: the event may have occurred at any point along the LOR. Time of Flight (TOF) techniques attempt to localise the annihilation event along the LOR using the difference in photon arrival times at the detectors [47], [59]. The difference in photon arrival times must be measured accurately, and therefore TOF-PET requires better timing resolution than non-TOF PET. Fast crystals, such as LSO, and faster electronics have enabled TOF to be used clinically in recent years. Figure 1.7 illustrates the principle of TOF-PET.

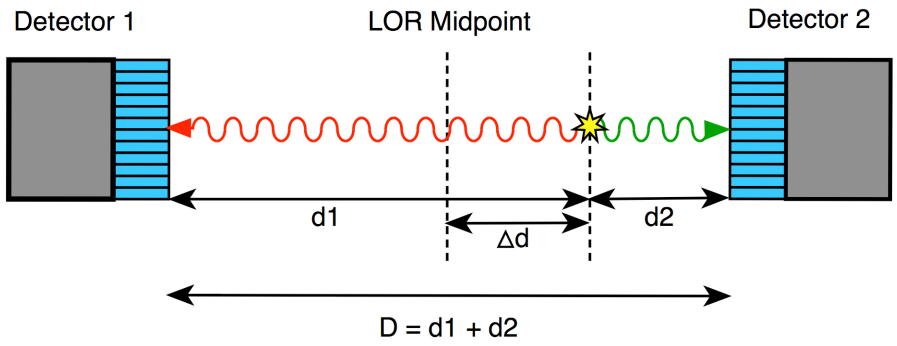


Figure 1.7: Principles of Time of Flight

An annihilation event originates at a distance *Δd* from the midpoint between two detectors, which are separated by distance *D (d1 + d2)*. Photon 1 (red) travels distance *d1* before being detected by Detector 1 and Photon 2 (green) travels distance *d2* before being detected by Detector 2. Photon 1 therefore travels *2Δd* further than Photon 2. If the difference in arrival times (*Δt*) can be measured, the distance *Δd* can be calculated:

$$\Delta d = \frac{c\Delta t}{2} \qquad \text{Equation 1.2}$$

With a fast enough detection system, TOF-PET could theoretically localise annihilation events to within a single image voxel: timing resolution of 10ps could isolate events to within a 3mm voxel. However, limitations in timing resolution introduce blurring in the estimation of *Δd* [59], and hence uncertainty of the position of the annihilation event. Modern LSO crystal-based PET scanners are currently capable of timing resolutions of between 500ps and 600ps [82], [83], giving a positional uncertainty of between 7.5cm and 9cm.

TOF restricts the backprojection of an acquired event to a small segment of the LOR, instead of blurring the event over the entire LOR, as illustrated in Figure 1.8. This reduces the statistical noise in the reconstructed image if the LOR segment is shorter than the size of the emission source [27]. The benefits of TOF-PET are increased Signal to Noise Ratios

(SNR), lower random coincidence rates and the ability to handle higher count rates [84]. TOF is of particular benefit when imaging larger patients, whose increased attenuation properties adversely affect image quality when conventional PET is used. TOF information can be used to improve image quality with standard acquisition times, or to achieve the same image quality with reduced acquisition times [85].

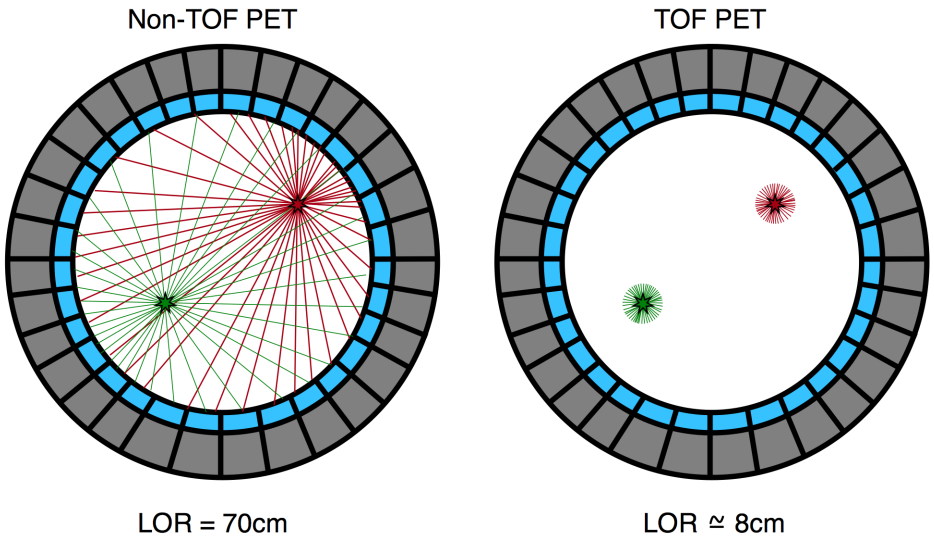


Figure 1.8: Positional uncertainty improved by TOF

The use of TOF means timing information is applied to each correction step within the iterative reconstruction loop (e.g. normalisation, randoms, deadtime, scatter, attenuation). As a result, TOF sinogram datasets are approximately 60 times larger than those containing non-TOF data [47]. The use of TOF data in clinical PET therefore requires significant storage space and processing power compared to non-TOF PET.

1.7.2 Point Spread Function Modelling

Techniques to correct for limitations in spatial resolution arising from the emission/detection process have been developed in recent years. A model of the system's PSF can be used by the reconstruction algorithm for such corrections. This requires knowledge of the PSF at every point throughout the FOV, as PSF is spatially dependent. The PSF model can be constructed using analytical calculations, Monte Carlo simulations or experimental measurement using point sources [19], [49]. The model can be applied in image space [86], [87], but is more commonly applied in sinogram space [49], [88] by incorporating PSF information into the system matrix. Such corrections should improve the uniformity of spatial resolution throughout the FOV [89] and reduce partial volume effects [37].

A PSF system matrix relates each voxel to more LORs than a non-PSF system matrix. Consequently, the reconstruction problem becomes even more ill-posed and requires more

iterations to reach convergence [90] [91]. Furthermore, neighbouring voxels in PSF-corrected images demonstrate greater correlations than uncorrected images, which changes the noise texture of the images [49], [92], [93]. At early iterations the benefit of PSF modelling is noise reduction, and not increased resolution [94].

A significant disadvantage of PSF correction is the introduction of artefactual edge enhancements known as Gibbs artefacts, visualised at the borders of tissues with large differences in intensity. Such artefacts may be caused by imperfections in the system matrix [95] and/or recovery of frequencies attenuated by the initial blurring process [71]. High frequency data lost in the initial detector blurring process cannot be recovered by PSF correction; however, the amplitude of the retained frequency content can be corrected. This produces a 'deblurred' PSF, which is closer to the ideal point source response, and improves the resolution of the reconstructed images. However, a steep cut-off in the frequency domain creates oscillating tails (ringing artefacts) in the spatial domain, as shown in Figure 1.9.

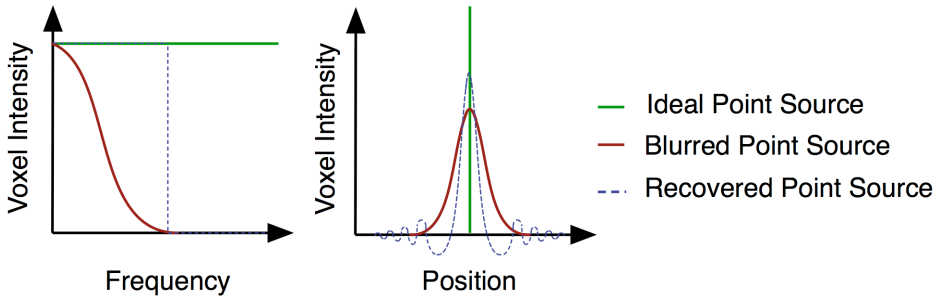


Figure 1.9: PSF correction with steep frequency domain cut-off

The effect of Gibbs artefacts is dependent upon the size of the lesion being considered. Enhanced edges may cause lesions to have apparent reduced activity at their centre. Sufficiently small lesions may have overestimated activity concentrations, caused by the summation of edge artefacts from opposing lesion boundaries [37]. Post-reconstruction filtering can mitigate these artefacts; however, care must be taken when selecting the filter width, as excessive smoothing will negate any advantage gained from PSF modelling.

1.8 Optimisation of PET Reconstruction

The West of Scotland PET Centre in Glasgow opened in 2007 with a General Electric Medical Systems (GEMS) Discovery STE PET-CT system. This system used BGO crystals, and therefore had no TOF capability. The reconstruction software included OSEM iterative reconstruction, although PSF modelling was not available clinically at that time. In 2011, a second PET-CT system was installed in Glasgow: the GEMS Discovery 690. This new

system used lutetium-based crystals instead of BGO, and the reconstruction software had optional TOF capability. Furthermore, the Discovery 690 OSEM reconstruction algorithm had the ability to apply PSF corrections, if desired. This new PET-CT system therefore presented Glasgow with significant advancements in technology. GEMS suggested both TOF and PSF corrections should be applied for general whole-body ^{18}F-FDG oncology imaging, in combination with a specific post-reconstruction filtering strategy, matrix size and number of effective iterations. However, the reconstruction parameters suggested by GEMS had not been widely accepted by the PET-CT imaging community and lacked justification through independent peer-reviewed studies. The use of PSF in particular was controversial as it was known to overestimate the activity in small lesions. There was, therefore, a requirement for the West of Scotland PET Centre to assess the effects of the newly available technology in order to select the optimum strategy for image reconstruction.

OSEM iterative reconstruction techniques are widely used in clinical PET practice. However, many thousands of different combinations of the selectable reconstruction parameters are possible, which makes their effects on reconstructed images challenging to assess. The optimal combination of such reconstruction parameters is dependent upon the clinical task being undertaken [96]: for example, the detection of small lesions for staging disease is likely to require a different combination of reconstruction parameters to the accurate quantification of lesion activity to assess treatment response. Furthermore, different areas of the human body have different attenuation and scatter properties, as well as variable ^{18}F-FDG physiological uptake: abdominal imaging presents different challenges to head and neck imaging. Image optimisation is, therefore, a significant challenge. Furthermore, whilst there have been many publications assessing the effects of various reconstruction parameters, there is no standardised methodology accepted by the PET-CT community for the assessment and optimisation of reconstruction parameters.

1.9 Liver Imaging in Cancer Patient Management

The liver is the most common site of haematogenous metastatic spread due to its rich blood supply [97], [98], most commonly from colorectal cancer (CRC), but also from cancers elsewhere in the GI tract, breast, lung and lymphomas. Globally, CRC is the third most common type of cancer and the fourth most common cause of cancer-related death [99]. 1.4 million new cases were diagnosed globally in 2012, which is projected to rise to 2.2 million new cases by the year 2030 [100]. Approximately 50-60% of CRC patients will develop liver metastases at some point in their disease [97]. Early diagnosis of liver metastases is crucial to effective patient management, as aggressive resection of liver metastases has been shown to improve survival. The only curative treatment for CRC liver

metastases is surgery, which can increase 5-year survival rates by up to 50% [101]. However, only approximately 20% of these patients will be suitable for resection (typically when patients have a solitary liver lesion, or a small number of lesions confined to a single lobe [102]). Incomplete resection does not improve patient survival, so knowledge of the full extent of liver disease is of vital importance when determining the most appropriate course of patient management [103]. Patients who are not suitable for resection may benefit from other treatment strategies, such as chemoembolization, radiofrequency ablation or system chemotherapy [104]. The ability to detect small, low intensity lesions is therefore critical [8].

MRI is well established as the gold standard for evaluating hepatic lesions [98]. However, the value of PET-CT in the management of hepatic metastases has also been recognised. A study by Yang et al [105] demonstrated that although MRI has superior spatial resolution, ^{18}F-FDG PET and MRI had similar performance in terms of liver lesion detection. A study by Kinkel et al [106] demonstrated that ^{18}F-FDG PET was, in fact, more sensitive than ultrasound, CT or MRI for detecting liver metastases from colorectal, gastric and oesophageal cancer. PET imaging currently plays an important role in determining a patient's eligibility for surgery (for example, to determine the number and resectability of lesions and to rule out extra-hepatic disease) and has been shown to influence patient management by multiple studies [107], [108].

Normal liver tissue tends to have a uniformly mottled appearance in ^{18}F-FDG imaging [109], with mild to moderate uptake (typical SUV_{mean} 1.3 – 3.0, SUV_{max} 3.0 – 4.0) [14]. Increased ^{18}F-FDG uptake in focal liver lesions may therefore be difficult to distinguish from healthy liver tissue due to this relatively high background activity. The detection of small lesions within the liver therefore requires a well-chosen balance between spatial resolution and image noise. The clinical importance of liver lesion detection, combined with the challenging nature of liver image interpretation, provided the motivation for selecting liver lesion detection as the clinical focus of the reconstruction optimisation project undertaken in this thesis.

1.10 Aims of Thesis

The overall aim of this thesis is to establish techniques which will facilitate image optimisation whilst considering clinical image context.

The primary aim is to develop a generic methodology to assess and optimise PET image reconstruction that can be applied to any clinical application.

The secondary aim is to optimise image reconstruction using the GEMS Discovery 690 PET-CT system, with respect to the detection of small liver lesions in ^{18}F-FDG-PET oncology imaging.

1.11 Organisation of Thesis

Chapter 2 reviews the literature on PET reconstruction optimisation and demonstrates the lack of consensus in the field. This chapter highlights that the optimum choices of reconstruction parameters for different clinical scenarios have yet to be established. Furthermore, this chapter highlights that there is no universally accepted approach to optimising PET reconstruction for a given clinical application.

Chapter 3 outlines the materials and methods used throughout this thesis. This describes the GEMS Discovery 690 system and the phantoms used to assess reconstruction image quality. The methods of image analysis common to many of the subsequent chapters are also described here.

Chapters 4 – 11 outline the practical work undertaken for this thesis. Each of these chapters begins with short background section and literature review specifically focussing on the topics being addressed in that chapter.

Chapter 4 investigates the effects overlapping image frames have upon image quality and justifies the choice of overlap size used for subsequent phantom and patient acquisitions. This chapter also proposes an amendment to European guidelines to account for the relationship between image noise and slice overlap when calculating minimum ^{18}F-FDG injection activities. Chapter 5 assesses how different combinations of reconstruction parameters affect the quantitative accuracy of reconstructed voxels and aims to validate the choice of reconstruction parameters recommended by GEMS for WCC. Chapter 6 assesses how different combinations of reconstruction parameters affect different metrics of image noise and presents a novel finding regarding the GEMS post reconstruction Gaussian filter implementation.

Chapter 7 develops a methodology for assessing spatial resolution in a clinically relevant manner. Chapter 8 then uses this methodology to assess how different combinations of reconstruction parameters affect spatial resolution.

Chapter 9 uses a phantom containing simulated lesions to assess how different combinations of reconstruction parameters affect lesion recovery. The experimental set-up

was designed to replicate the activity concentrations typically measured within patient livers. The findings of this chapter, combined with those of the preceding chapters, were used to select two reconstruction strategies which merited assessment using patient data. This chapter also recommends the use of specific image quality metrics to predict human observer preferences for lesion detection. Chapter 10 then compares the lesion phantom to patients with varying body sizes to predict which patients the recommended reconstructions will be most applicable to.

An observer study using patient data is performed in Chapter 11. Experienced PET reporting clinicians were involved in image assessment which compared different reconstruction strategies. Image quality, reporting confidence and lesion detection are all evaluated. The results of this study are used to recommend a particular reconstruction strategy to optimise small liver lesion detection.

Finally, Chapter 12 discusses the major findings of the thesis and draws relevant conclusions. It presents a generic methodology for assessing and optimising reconstruction techniques, which can be followed for any clinical scenario. Furthermore, it presents an optimised reconstruction strategy for liver lesion detection. Chapter 12 also outlines further work to be undertaken in the continuation of this research.

Chapter 2 : Literature Review

Image reconstruction parameters affect the noise, spatial resolution, lesion-to-background contrast and quantitative accuracy of PET images. This in turn affects the ability of an observer to detect true lesions whilst minimising false positives. Convergence can vary as a function of patient size, lesion size and uptake, and image reconstruction algorithm. Theoretically, each lesion within a patient could have a different optimal reconstruction; however, clinical reconstruction strategies often use a fixed reconstruction method that provides reasonable images over a range of imaging scenarios [110], [111]. Many studies have investigated the effects of various reconstruction parameters upon image quality: however, there is no consensus in the literature as to which reconstruction parameters optimise lesion detection.

This chapter provides an overview of the major themes and discusses publications of significant relevance to this thesis. Subsequent experimental chapters include further literature reviews that specifically target the particular topic being addressed.

2.1 International Guidance for ^{18}F-FDG PET Reconstruction

The European Association of Nuclear Medicine (EANM) published updated guidelines for ^{18}F-FDG tumour imaging in 2015 [14], which have also been adopted by the United States' Society of Nuclear Medicine and Molecular Imaging (SNMMI). These guidelines aim to standardise image quality for oncology between centres taking part in multicentre trials, rather than optimise reconstruction on specific PET-CT systems for specific clinical tasks. The guidelines make the following image reconstruction recommendations:

- Reconstructed voxel sizes should be within 3.0 – 4.0mm in any direction
- TOF should be used, when available
- PSF may be applied, but may require additional filtering for multicentre studies
- The FWHM of post reconstruction filters should not exceed 7mm

No specific guidance is included for z-axis filtering or for the number of reconstruction iterations.

These recommended reconstruction parameters originate in part from a 2013 report by the Quantitative Imaging Biomarkers Alliance (QIBA) [112], which stated that voxels sizes should be between 3mm and 4mm in all three dimensions for whole body imaging. However, a subsequent QIBA report in 2014 [113] stated that voxels of between 2mm and 3mm in all three dimensions should be achievable with current technology (as long as the smaller voxels do not introduce artefacts). The 2014 QIBA voxel size recommendation was not adopted by the 2015 EANM guidelines.

To the author's knowledge, there are no other published guidelines which specify reconstruction parameters to be used for ^{18}F-FDG tumour imaging. In particular, there are no accepted guidelines for liver imaging or for maximising small lesion detection.

2.1.1 Accreditation for Multi-Centre Clinical Trials

Visual assessment of PET images may be sufficient for staging and restaging patients [114]–[116]. However, the accuracy of visual interpretation is user-dependent, and therefore difficult to use in multi-centre clinical trials. Evaluation of solid tumour response to therapy is challenging and requires some form of quantification [114]. Diffuse increased liver uptake, which could represent a variety of conditions, could be overlooked when relying on visual interpretation only [117]. Quantitative analysis provides objective and more accurate measurements with less user-dependence than visual assessment alone. Applying such analyses to baseline and early response-to-treatment scans can potentially identify responding and non-responding patients more accurately and improve the prognostic value of early scans [118]. SUVs are the most commonly used semi-quantitative parameter for analysis of ^{18}F-FDG uptake [14]. Current recommendations are that tumour SUVs should be reported, and that sources of SUV bias and variance should be well understood and controlled [119]. The EANM tumour imaging guidelines recommend that SUV_{max} is reported for each lesion, when deemed clinically relevant [14]. However, other metrics such as total lesion glycolysis (TLG) and metabolic tumour volume (MTV) are now being increasingly used [120].

EANM guidelines refer to EANM Research Ltd (EARL) guidelines for scanner accreditation, which are followed by imaging centres that wish to participate in multi-centre clinical trials [121], [122]. The EARL guidelines outline a specific quality control phantom acquisition using a National Electrical Manufacturers Association (NEMA) International Electrotechnical Commission (IEC) Body Phantom. This phantom is a fillable torso-sized shell containing six fillable spheres of varying diameters and is widely used to assess PET image quality. This phantom is discussed in more detail in Chapter 3. Centres seeking EARL accreditation must ensure their image reconstruction methodology produces images that meet EARL specifications in terms of SUV recovery and noise measurements.

Other accreditation programmes include those of the National Cancer Research Institute (NCRI) in the UK and the SNMMI Clinical Trials Network in the United States. NCRI accreditation closely follows EARL procedures, while SNMMI accreditation uses a novel anthropomorphic chest phantom [123]. Glasgow's Discovery 690 PET-CT system successfully received accreditation from both NCRI and SNMMI; however, PSF corrections had to be switched off to pass both accreditations. NCRI accreditation initially failed when PSF was applied, as the recovered SUV_{max} of the spheres exceeded the maximum limits; this was particularly true for the smaller spheres (it should be noted that the acceptable range of results for NCRI are within two standard deviations of the 'average' accredited scanner result). SNMMI requested PSF to be switched off at the outset.

Accreditation programmes assess areas of background activity as well as lesion recovery. Both SNMMI and NCRI measure SUV_{mean} in an area of uniform background activity within the phantom. The scanner's calibrations are acceptable if the measured SUV_{mean} equals $1.0 \pm 10\%$ (i.e. within 10% of the expected SUV_{mean} value when radioactivity is uniformly distributed throughout the object in the FOV). This is a widely accepted technique for assessing the quantitative accuracy of background voxel values [124], but it does not quantify noise within the image. EARL accreditation assesses image noise by measuring the coefficient of variation (COV) of background voxel values. Their upper COV limit for acceptable image noise is 15% (when background activity concentration is chosen to be similar to that of a typical patient). However, this singular method of noise measurement does not fully assess the noise characteristics of PET images, as will be discussed.

Guidelines and publications focussing on multi-centre trials prioritise consistency amongst participating centres over the optimisation of any particular PET-CT system for any particular clinical application. Centres which adopt new technology as it emerges may find themselves unable to fully utilise these new techniques for clinical trial patients if their optimised reconstructions, though improved, are inconsistent with current trial standards.

As new technologies become more widely adopted (following thorough evaluation in peer-reviewed literature), accreditation programmes should re-evaluate their standards to account for such advancements. The work undertaken in this thesis will contribute to the evidence-based evaluation of reconstruction methods.

2.2 Assessment of PET System Performance

All newly installed PET systems undergo a set of acceptance testing procedures. These procedures evaluate different aspects of the system's performance and compare the results against the manufacturer's specifications. Acceptance tests are typically repeated as part of the system's routine quality control procedures [125]; for example, manufacturers may recommend repeating these tests on an annual basis. Standard procedures for acceptance testing are outlined by NEMA documentation [70] and use a specific set of phantoms. Tests include spatial resolution, scatter fraction, sensitivity and image quality. It should be noted that while the 2012 version of the NEMA acceptance testing standards applies to this thesis, the standards were updated in 2018.

Whilst NEMA style tests are useful for comparing an individual system's performance against its specifications, and for comparing systems produced by different manufacturers, many aspects of the tests do not reflect clinical imaging conditions. For example, the NEMA test for spatial resolution uses point sources in air, which are reconstructed using filtered back projection. Typical NEMA point source FWHM results are in the order of between 3mm and 5mm; however, this methodology does not represent a clinical scenario.

A similar disconnect between NEMA acceptance testing protocols and clinically realistic imaging conditions exists for conventional gamma cameras [126]. An attempt was made to address this by the Gamma Camera Assessment Team (GCAT), established by the Medical Devices Agency (MDA) [127], [128]. GCAT evaluated and compared different gamma camera systems using a more clinically-orientated protocol; for example, scatter media was used in their assessments of spatial resolution. To the author's knowledge, there has been no comparable attempt to standardise the assessment of PET system performance in a clinically relevant manner.

2.3 Optimisation of PET Reconstruction Parameters

Many studies have investigated the effects of various combinations of reconstruction parameters upon different aspects of image quality. Qualitative and quantitative analysis has been undertaken on both phantom studies and patient studies. Although there are some

commonly used methodologies in the literature, as shall be demonstrated in this chapter, there is no universally accepted approach to assessing and optimising reconstruction parameters for a particular clinical task.

Various image metrics have been used by quantitative assessments in the literature. Phantom studies with known sphere and background activity concentrations often use recovery coefficients to assess image convergence and the accuracy of voxel values within lesions (100% recovery indicates voxel values exactly match the known activity concentration). Image noise is typically assessed using ROIs in regions of background activity to calculate either voxel-to-voxel noise metrics (such as COV or standard deviation) or region-to-region noise metrics (such as background variability). Lesion detection is often assessed using contrast and SNR, as well as observer studies using both human or mathematical observers. SUVs are frequently used in both phantom and patient studies. The metrics used in this thesis are discussed in more detail in the following chapter.

2.3.1 TOF, PSF and Effective Iterations

The effects of TOF and PSF corrections have been assessed by many studies. As these corrections affect image convergence, they are often studied in conjunction with the number of effective iterations.

Several studies have demonstrated TOF reconstructions require fewer effective iterations to reach convergence than non-TOF reconstructions. For example, a study by Bettinardi et al using the Discovery 690 [19] found that TOF required between 54 and 90 effective iterations to reach convergence, while the non-TOF reconstructions typically required > 90 effective iterations. This study also demonstrated that lesions in TOF reconstructions converged to more accurate values than the non-TOF reconstructions. Studies by Jakoby et al [111], Surti et al [129], Karp et al [130], Kadrmas et al [131], Lois et al [18], Akamatsu et al [132] and Schaefferkoetter et al [8], using a variety of PET-CT systems, all demonstrated similar results when TOF was applied with increasing iterations.

Karp et al [130] and Lois et al [18] noted that, for the same number of effective iterations, the voxel-to-voxel noise in the TOF reconstructions was greater than that of the non-TOF reconstructions. This is as a result of TOF's faster convergence. Wilson and Turkington [133] advise that when comparing TOF and non-TOF image quality, the number of effective iterations is a poor choice to equalise. However, even when effective iterations are equalised, studies show the increase in noise as a result of applying TOF is typically offset by the improvement in the lesion signal, leading to improved lesion signal-to-noise ratios [18], [43], [134].

Several studies report that TOF is of particular advantage when imaging large objects. Bettinardi's study [19] compared large (27cm x 39cm x 30cm) and small (24cm x 30cm x 24cm) torso phantoms and observed that TOF produced a greater improvement in lesion detection in the larger phantom compared to the smaller phantom. A patient study by Surti et al [110] found that TOF imaging led to improvements in both liver and lung lesion detection, and the extent of the improvement increased as the patient size increased. Similar findings were observed by Jakoby et al [111], Karp et al [130], Lois et al [18] and Schaefferkoetter et al [8]. However, a phantom study by Wilson and Turkington [133] assessed the impact of TOF on the SNR of small spheres in a range of phantom sizes and concluded that no TOF image quality improvement is expected for activity diameters of ≤ 17.5cm. Prieto et al [135] and Bettinardi et al [19] assessed images of a brain phantom (18cm diameter) and found that TOF reconstruction results were similar to those of non-TOF reconstructions, both quantitatively and qualitatively.

In contrast to TOF, PSF modelling has been shown to increase the number of effective iterations required to reach convergence. Studies by Rapisarda et al [86], Akamatsu et al [132], [134], Bettinardi et al [19] and Kenny and McGowan [136] demonstrated PSF algorithms require high numbers of effective iterations to reach convergence (e.g. >180 effective iterations on the Discovery 690). At these high numbers of effective iterations, the recovery of hot lesions exceeded that of the TOF-only reconstruction; however, some studies noted the time taken to perform the required effective iterations may be impractical in the clinical setting. Patient studies by Andersen et al [137], Kawashima et al [138] and Sheikhbahaei et al [139] did not investigate the effect of varying effective iterations, but did demonstrate PSF increased lesion SUVs (both mean and maximum).

Several studies have assessed qualitatively the noise in PSF reconstructions and concluded the noise has a more 'correlated' appearance than non-PSF reconstructions [19], [49], [92], [93]. Noise has been assessed quantitatively by several studies, and the conclusions drawn depend upon the method of noise measurement. Akamatsu et al [134], Prieto et al [135], Rapisarda et al [86] and Panin et al [140] all used voxel-to-voxel measurements and concluded that PSF improves image noise. Sureau et al [87] noted that PSF increased correlations between adjacent voxels, and therefore increased spatial homogeneity within uniform background regions. A study by Tong et al [92] found that PSF reduced noise when measured by voxel-to-voxel techniques; however, PSF did not improve noise measured using region-to-region techniques. Rahmim and Tang [93] state that the use of voxel-to-voxel noise measurements may overstate image improvements as a result of PSF: for example, human observers may not agree that correlated background activity improves image quality.

Studies have also shown that PSF has led to increased recovery of small objects. Lasnon et al [141] noted that the most pronounced PSF advantage in clinical scans was on the quantitative values of sub-centimetre nodes. Bellevre et al [142] assessed the use of PSF in staging the axilla in breast cancer patients and found that PSF performed better than non-PSF in detecting nodes < 7mm. Rapisarda et al [86] and Sureau et al [87] found that PSF increased contrast recovery, especially for the smallest lesions, when compared to non-PSF. Kenny and McGowan [136] assessed PSF on a Discovery 690 using a NEMA phantom and observed significant increases in contrast recovery for the smaller spheres. They also observed small lesion SUV_{max} increases of up to 78% in patient data.

However, there are concerns about PSF edge artefacts that may cause over/undershoots in voxel values near sharp boundaries in the reconstructed images. These artefacts can cause hot structures within an image to have enhanced edges and colder centres. Nagaki et al [143], Bettinardi et al [19] and Rapisarda et al [86] all observed overemphasis of voxel values at high contrast boundaries when PSF was applied. When the hot structure/lesion is sufficiently small, edge artefacts can merge which can lead to overshoots in voxel values [37]. This may cast doubt on some clinical studies that claim PSF improved small lesion contrast recovery; some of the improvement may in fact be artefactual. A phantom study by Kidera et al [144] found that in lesion spheres smaller than 17mm diameter, edge artefacts appeared as a sharp peak at the centre of the sphere, resulting in overestimation of the sphere radioactivity. Spheres of 22mm diameter or greater demonstrated the artefact as edge enhancement. These artefacts increased as the number of iterations increased and as the sphere-to-background ratio increased.

Several studies have evaluated both PSF and TOF together and compared the results to those obtained using each correction on its own, or when neither correction is used. Bettinardi et al [19] stated that TOF improves SNR and acts as a convergence ‘accelerator’ which allows PSF to recover a better signal at lower iterations and reduce the image noise. Schaefferkoetter et al [8] concluded applying TOF and PSF together was particularly beneficial when imaging low intensity lesions in larger patients. Taniguchi et al [145] concluded that PSF and TOF together produced the best SNR and contrast recovery results. Akamatsu et al [134], [146] Kadrmas et al [131] and Matheoud et al [147] found that the combination of PSF and TOF provided better image quality than using either PSF or TOF alone, in both phantom and patient studies.

Rogasch et al [148] concluded that applying PSF and TOF together significantly improved spatial resolution; however, they also stated that the effects of PSF edge effects require further investigation as it caused substantial overestimations of SUV_{max}. Kuhnert [149] recommended that two reconstructions should be performed in routine clinical practice:

TOF and PSF should be used together to produce a data set for optimum visual assessment, while a second data set should be used for quantification (e.g. additional filtering applied to the TOF + PSF reconstruction).

PSF corrections have therefore not been universally applied in routine clinical imaging, particularly when quantification is important. Kenny and McGowan [136] concluded that PSF should be used with caution when comparing longitudinal studies; later publications by this group confirmed they did not use PSF in their standard of care GEMS Discovery 690 reconstructions [150], [151]. A study by Nakamura et al [152] stated that PSF reconstructions are not appropriate for SUV evaluation, as small lesion SUVs were overestimated by up to 50%. Alessio and Rahmim [153] state that some imaging sites, including theirs, do not use PSF when performing quantitative pharmacokinetic imaging. Munk et al [154] concluded it is safer not to use PSF for quantitative purposes, as they found the expected monotonic relationship between sphere size and recovery coefficients was broken in an unreproducible manner.

2.3.2 Post Reconstruction Filters

GEMS PET-CT systems typically employ two independent methods of filtering: a Gaussian filter that acts upon voxels within each transaxial slice and an axial filter that acts upon voxels in adjacent transaxial slices (also referred to as a ‘z-axis filter’). The use of a Gaussian transaxial filter is common to most commercial PET-CT systems; however, the use of an independent z-axis filter hasn’t been universally adopted by all vendors. As a result, many studies have assessed the effects of post-reconstruction transaxial filters in combination with PSF and TOF techniques, while few publications have even mentioned z-axis filtering in their methodologies.

Most studies assessing GEMS PET-CT reconstruction methods do not include any details of the z-axis filtering strategy. Studies by Bettinardi et al [19] and Kawashima at al [138] assessed the effects of altering other reconstruction parameters but simply stated a ‘standard’ z-axis filter was used for all reconstructions. De Ponti et al [155] stated a ‘standard’ z-axis filter was used in their NEMA image quality assessment of the GEMS Discovery 600 PET-CT system. Wilson and Turkington [133] state that z-axis smoothing was applied in their comparison of TOF and non-TOF images, but no more details are provided. The GEMS White Paper on their implementation of PSF [89] includes examples of whole body patient images acquired using both ‘light’ and ‘standard’ z-axis filtering; however, the effects of these filters are not assessed in this paper. To the author’s knowledge, there are no publications which specifically assess the effects of different z-axis filtering strategies.

Kadrmas et al [131] investigated the use of different post reconstruction transaxial Gaussian filter widths with TOF and PSF in whole body imaging. They concluded that little or no filtering should be used with PSF reconstructions in order to maximise lesion detection, as voxel correlations introduced by PSF produce a similar effect to a smoothing filter. They concluded a filter with FWHM ≈ 5mm achieved the best results for plain OSEM and TOF reconstructions. This filter strategy was also adopted by Bellevre et al [142] and Sheikhbahaei [139].

Akamatsu et al [134] also assessed different Gaussian filter FWHMs (0mm → 10mm) with TOF and PSF in whole body imaging. They found that PSF reconstructions achieved their optimal SNR when <2mm FWHM filters were used, whilst non-PSF required a filter FWHM between 4mm and 6mm. Subsequent studies by this group [132], [146] used a 2mm FWHM Gaussian filter for PSF reconstructions and a 4mm FWHM Gaussian filter for non-PSF reconstructions. This filter strategy was also adopted by Andersen et al [137] and Taniguchi et al [145]. Schaefferkoetter [8] also looked at different filter FWHM in conjunction with TOF and PSF in whole body imaging. Whilst most studies recommend a choice of two filtering strategies based upon whether or not PSF is applied, this study recommended three filtering strategies to maximise SNR: 6mm FWHM filters for plain OSEM and TOF, 4mm FWHM filter for PSF-only and no filtering at all when both TOF and PSF are applied.

In contrast to the limited PSF filtering strategies of the above studies, Panin et al [88] used a 7mm filter for PSF reconstructions to produce ‘clinically acceptable’ whole body images. However, they also suggested unsmoothed PSF reconstruction may provide more precise structural information. A study by Lasnon et al [67] also suggests a 7mm FWHM filter can be used on PSF images to mitigate Gibbs effects and produce recovery coefficients comparable to non-PSF reconstructions. Munk et al [154] found combining PSF with a post-reconstruction filter of 3-4mm could restore the expected monotonic relationship between sphere size and recovery coefficient and still maintain some of the PSF recovery improvements.

2.3.3 Voxel Sizes

Current practice for whole body FDG scanning uses voxels of approximately $4mm^3$ [14], [156]–[158], which is of the order of the NEMA spatial resolution [159]. However, smaller voxels can be achieved by increasing the matrix size and/or reducing the transaxial FOV. It should be noted that the GEMS Discovery 690 has a fixed z-axis voxel dimension of 3.34mm, due to the physical configuration of the system.

A phantom study by Morey et al [157] compared two different voxel sizes: $4mm^3$ and $2mm^3$. Although background noise was increased as a result of using smaller voxels, they concluded that $2mm^3$ voxels significantly improved lesion detection performance for focal lesions in noisy backgrounds. Furthermore, they demonstrated the smaller voxels improved lesion detection in whole body scanning to a greater extent than the application of PSF. They also noted that the use of the smaller voxels significantly increased reconstruction time and storage requirements.

Adams et al [66] assessed the effects of different voxel sizes and filter FWHMs upon small object SUV_{mean} in a phantom using a Discovery 690. Adjusting both the matrix size and transaxial FOV created the following transaxial plane voxel dimensions: $5.47mm^2$, $3.90mm^2$, $3.65mm^2$ and $2.73mm^2$. Gaussian Filter FWHMs were varied between 0mm and 5mm. The smallest voxels combined with no Gaussian filtering were shown to produce the most accurate SUV_{mean} results, as the smaller voxels are more likely to sample the peak of the lesion. This study did not discuss the impact of voxel size on image noise.

A study by Zhang et al [160] using both phantoms and patients compared three different voxel sizes: $1mm^3$, $2mm^3$ and $4mm^3$. They found that using smaller voxels improved lesion detection and contrast recovery; however, it should be noted that this study was conducted on a solid-state PET detection system capable of greater spatial resolution than traditional PMT-based detection systems. This study did not specifically discuss the impact of voxel size on image noise.

Koopman et al [158] compared $4mm^3$ voxels with $2mm^3$ voxels using phantoms and patients. Qualitative and quantitative analysis of the phantoms demonstrated that, although smaller voxels produced higher background noise, they also improved small lesion detection, contrast recovery and SNR for lesions smaller than 2cm. The study noted voxel size had little effect on lesions of 2cm or greater.

2.4 ^{18}F-FDG Liver Imaging

2.4.1 Healthy Liver as an Internal Reference

Normal liver tends to have a uniformly mottled appearance in ^{18}F-FDG imaging [109], with mild to moderate uptake (typical SUV_{mean} 1.3 – 3.0, SUV_{max} 3.0 – 4.0) [14]. As the uptake is relatively uniform, clinical image quality is often assessed by placing ROIs over the liver to measure voxel-to-voxel noise or SNR [6], [161]. Liver uptake can also be used as an internal reference uptake for comparison with lesion uptake, providing a semi-quantitative estimation of tumour metabolism [162]. The Deauville 5-point scale is used to assess

lymphoma scans by comparing the most active lesion uptake with liver uptake [163]. Positron Emission tomography Response Criteria In Solid Tumours (PERCIST) also use liver uptake as reference for comparison with lesion uptake [6]. Alternative reference sites must be used, however, when the liver contains widespread abnormalities; for example, blood pool, mediastinum and cerebellum.

2.4.2 Liver Reconstruction Studies

Many studies have assessed the performance of different reconstruction strategies by measuring noise within healthy liver images. Studies by Yan et al [164] and Taniguchi et al [145] investigated the effects of TOF and PSF in patient data and measured SNR and COV respectively in liver tissue. Both studies concluded using PSF and TOF together produced the best noise results. Akamatsu et al [134] also measured liver SNR to assess the effects of PSF and TOF, concluding optimum results were obtained using both corrections together. A later study by the same group [146] reached the same conclusion when using liver COV to assess the use of PSF and TOF in patients of varying weight. Armstrong et al [165] used liver SNR to demonstrate TOF could be used to reduce scan times and/or injected activities without compromising image quality.

Fewer studies have assessed the effects of reconstruction parameters specifically upon liver lesion detection. Surti et al [110] added 1cm diameter spherical lesions to healthy patient liver data by combining patient and lesion-only sinogram datasets. This study used a Philips Gemini PET-CT system. Reconstructions were then performed with and without TOF (this study did not assess other reconstruction parameters). TOF was shown to improve lesion detection and localisation. A similar approach to simulating lesions in healthy liver data was taken by Schaefferkoetter et al [8] using a Siemens Biograph PET-CT system. This study concluded the combination of PSF and TOF produced optimum lesion detectability when used with no post-reconstruction filtering and 24 effective iterations. Rogasch et al [166] assessed the effects of TOF and PSF on liver lesion SUV_{max} results (other reconstruction parameters were not investigated). This study also used a Siemens Biograph PET-CT system. TOF was found to produce the greatest SUV_{max} in low contrast lesions (tumour to background ratio of <5:1), while PSF was shown to produce the greatest SUV_{max} in higher contrast lesions. The PSF effect was likely related to Gibbs artefacts, which are more likely to occur in higher contrast objects. The application of both PSF and TOF together increased SUV_{max} across all lesion-to-background contrast levels. To the author's knowledge, there has been no similar study conducted using the GEMS Discovery 690 PET-CT system.

2.5 Conclusions

This review has demonstrated that, although reconstruction guidelines and accreditation programmes exist to ensure consistent image quality in clinical trials, there is no established guidance for optimising image reconstruction for specific clinical tasks. Such guidance would be particularly useful when new technological advancements, such as TOF and PSF, become available. Additionally, there are no universally agreed methods for assessing different aspects of PET image quality in a clinically relevant manner.

The review has also demonstrated a lack of consensus regarding the application of PSF modelling techniques. Some studies concluded no post-reconstruction filtering is required when PSF is applied, whilst others recommended the use of filters to mitigate potential Gibbs artefacts. Some studies have recommended PSF should not be used at all, particularly for quantitative studies.

Furthermore, there is no established optimum reconstruction method for detecting small lesions within the liver in ^{18}F-FDG PET imaging. A small number of studies have assessed the use of both PSF and TOF in liver lesion detection and concluded applying TOF and PSF yields optimum results. However, none of these studies used the GEMS Discovery 690 PET-CT system.

This thesis therefore aims to develop a methodology to assess different aspects of PET image quality in a clinically relevant manner. These methods will be used to assess the effects of different combinations of reconstruction parameters upon image quality, with a view to optimising the detection of small lesions within the liver using the GEMS Discovery 690 PET-CT system.

Chapter 3 : Materials and Methods

This chapter describes the materials and methods which underpin most, if not all, of the practical chapters in this thesis. Each practical chapter has its own section that covers the materials and methods specific to each individual chapter.

3.1 GEMS Discovery 690 PET-CT System

The PET tomograph used in this investigation was the GEMS Discovery 690 PET-CT system, which combines a lutetium-yttrium orthosilicate (LYSO) PET tomograph with a 64-slice CT scanner. The system is shown in Figure 3.1.

Figure 3.1: GEMS Discovery 690 PET-CT system

3.1.1 Hardware

Figure 3.2 illustrates the crystal arrangement of the Discovery 690 PET system. Each individual LYSO crystal has dimensions 6.3mm (axial) x 4.2mm (transaxial) x 25mm (radial). 54 crystals in a 9x6 arrangement form a detection block. The dimensions of this square block of crystals are marginally increased by the use of reflective material between crystals to prevent optical spillover. The block of crystals is optically coupled to a four-anode PMT (Figure 3.2 (B)) to complete a detector block.

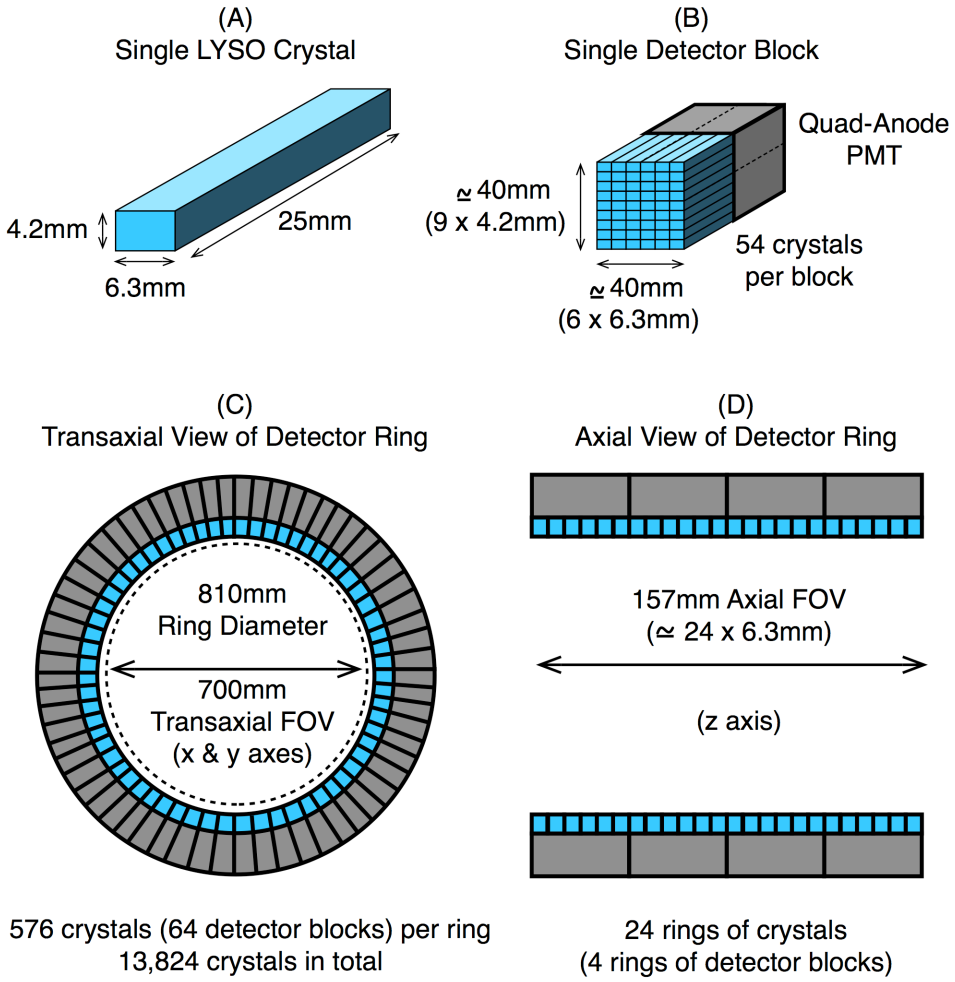


Figure 3.2: GEMS Discovery 690 crystal arrangement

The PET tomograph consists of four rings of 64 detector blocks (or 24 rings of 576 crystals). This creates a crystal ring diameter of 810mm (Figure 3.2 (C)), which provides an imaging FOV of 700mm [167]. The axial FOV created by the 24 crystal rings is 157mm (Figure 3.2 (D)). The PET tomograph has 256 detector blocks in total, containing 13,824 crystals. It operates only in 3D acquisition mode: all 24 crystal rings are able to form LORs with each other.

The Discovery 690 is equipped with a powerful processing system (*IBM BladeCentre*), designed to accelerate image reconstruction and data processing.

3.1.2 Reconstruction Software

3.1.2.1 OSEM Algorithm

The standard reconstruction method used by the Discovery 690 is a fully 3D OSEM reconstruction, known as *VUE Point HD* (VPHD). This algorithm includes corrections for scatter, randoms, and attenuation inside the iterative loop [36]. The VPHD algorithm uses a system matrix that accounts for the system's geometry (block-based crystal distribution and detector curvature), normalisation and dead time.

3.1.2.2 TOF and PSF

TOF data can be included in the reconstruction process by selecting the *VUE Point FX* (VPFX) algorithm. This applies timing information to each correction step within the iterative loop [47]. The Discovery 690 uses a timing kernel of 650ps in the reconstruction process [168]. Applying this to Equation 1.2 translates to a positional uncertainty of 9.75cm.

The reconstruction can also include PSF modelling by selecting the *SharpIR* option. This PSF model was developed by measuring the detector response to a point source placed at discrete locations throughout the radial and axial directions in the FOV [49], [89]. The detector response was then incorporated into the system matrix used by the OSEM reconstruction algorithm.

Table 3.1 summarises the four reconstruction methods used in this thesis. The abbreviations in the right-hand column are used throughout this thesis.

GEMS Notation	Description	Thesis Notation
VPHD	OSEM Algorithm	HD
VPHD-S	OSEM plus SharpIR	PSF
VPFX	OSEM plus TOF	TOF
VPFX-S	OSEM plus TOF and SharpIR	PSF+TOF

Table 3.1: Thesis notation for reconstruction methods

3.1.2.3 Effective Iterations

The OSEM algorithm allows the user to select the number of subsets and iterations used for reconstruction. The maximum permitted number of subsets is 48, whilst the maximum permitted iterations is 200; the maximum number of effective iterations is therefore 9,600.

In this thesis, the number of subsets was fixed at 18, while the number of iterations was varied. This follows the methodology used by Bettinardi et al [19] in their assessment of the Discovery 690's performance. Most experiments in this thesis used between 18 and 540 effective iterations. This range of effective iterations was chosen to provide a reasonable spread of results and should demonstrate convergence under most circumstances (this goes beyond the range used by Bettinardi et al, who stopped at 360 iterations). The remainder of this thesis describes reconstructions in terms of the effective iterations employed, instead of subsets and iterations.

3.1.2.4 Voxel sizes

The Discovery 690 has unevenly sampled projections due to the system geometry (block-based crystal distribution and detector curvature). The OSEM algorithm uses this projection data to reconstruct an image volume consisting of cuboid voxels. *Distance driven projectors* use the known detector boundaries and the position of voxels within the FOV to determine a detector's contribution to a particular voxel in the forward projection process, and a voxel's contribution to a particular detector during the back projection process [169].

The dimensions of the reconstructed voxels depend on both selected matrix size and transaxial FOV. Three matrix sizes can be selected for the transaxial images: 128*128, 192*192 and 256*256. The maximum possible transaxial FOV is 700mm, which is used for whole-body imaging. This can be reduced, e.g. for paediatrics or brain imaging, which in turn reduces the voxel dimensions in the x & y planes. The minimum possible FOV is 64mm; selectable voxel sizes therefore range from 0.25mm up to 5.47mm. The axial FOV is fixed at 157mm, and the axial sampling (z-axis voxel size) is fixed at 3.34mm. This produces 47 transaxial image slices for a single frame acquisition and cannot be altered by the user.

3.1.2.5 Post-Reconstruction Filtering

The Discovery 690 has two separate filter options that can be applied to the OSEM reconstructed data: one that is applied transaxially (x & y axes) and one that is applied axially (z-axis).

The transaxial filter is a two-dimensional Gaussian filter, which is defined by selecting the filter's FWHM in millimetres. As the filter is Gaussian, the standard deviation, σ, can be calculated for a given FWHM as follows:

$$FWHM = 2.35 * \sigma \qquad \text{Equation 3.1}$$

The GEMS filter design is truncated to ± 4σ, and therefore the full width of the filter is 8σ. As the truncation of the filter is minimal, it is assumed that the filter closely approximates a Gaussian curve in the frequency domain. The two-dimensional filter is implemented as a one-dimensional filter over the transaxial image rows and a one-dimensional filter over the transaxial image columns (Dr Charles Stearns, personal correspondence, 21st July 2014).

The z-axis filter is a 3-point average filter with four possible weight settings, as shown in Table 3.2. This filter is applied to the corresponding voxels in three contiguous transaxial image slices. The 'heavier' the filter weighting, the more smoothing is applied between slices.

Z-Axis Filter	Filter Coefficients
None	n/a
Light	[1 – 6 – 1] ÷ 8
Standard	[1 – 4 – 1] ÷ 6
Heavy	[1 – 2 – 1] ÷ 4

Table 3.2: GEMS Discovery 690 z-axis filter weights

3.1.2.6 Vendor Suggested Reconstruction Parameters

The following combination of reconstruction parameters was suggested by GEMS for whole-body ^{18}F-FDG imaging upon the system's installation:

- Both TOF and PSF applied (VPFX-S algorithm)
- OSEM reconstruction with 18 subsets and 3 iterations (54 effective iterations)
- 700mm FOV with 192*192 matrix (3.65mm voxel size in transaxial plane)
- Post-reconstruction Gaussian filter with 4mm FWHM
- Standard z-axis filter

The time taken to reconstruct a single frame acquisition using this combination of parameters is 1 minute and 46 seconds.

3.1.3 Quality Control and System Calibration

The Discovery 690 is subject to a robust quality control program to ensure consistent quantitative accuracy of reconstructed images over the system's life cycle. Detector performance is assessed daily by uniformly exposing the crystals to a long-lived ^{68}Ge rod source and comparing measurements against expected values. Weekly image quality tests are performed by acquiring, reconstructing and analysing images of a uniform phantom. WCC is performed on a quarterly basis, as advised by GEMS, to ensure accurate quantification in the reconstructed images. Annual quality control tests involve repeating the

acceptance tests performed at system installation and comparing with baseline values to ensure there is no degradation in system performance over time.

All experiments undertaken during this study were performed after verifying all appropriate quality assurance (QA) tests had been performed successfully, registration between corresponding PET and CT image volumes was satisfactory and WCC was up-to-date.

3.2 Phantoms

3.2.1 ^{68}Ge Cylindrical Phantom

Many clinical PET departments use a long-lived phantom to perform regular quality control procedures. As they are long-lived, these phantoms can be calibrated to national standards. ^{68}Ge has a half-life of 271 days and decays by electron capture in equilibrium with its short-lived daughter product ^{68}Ga (68 minute half-life), which is a positron emitter. Table 3.3 summarises the differences between ^{68}Ge/^{68}Ga and ^{18}F.

	^{68}Ge/^{68}Ga	^{18}F
Half-Life	271 days	109.5 minutes
Branching Factor	89%	97%
Maximum Positron Energy	1.90MeV	0.65MeV
Mean Positron Range in water [170]	2.24mm	0.64mm
Calibrated to National Standards	Yes	No

Table 3.3: Comparison between ^{68}Ge/^{68}Ga and ^{18}F

A uniform, cylindrical ^{68}Ge phantom (Figure 3.3) with 20cm diameter and 20cm length is used in Chapters 4, 5 and 6 of this thesis. The phantom's radioactivity concentration was calibrated to national standards with a stated uncertainty of ± 3% (the phantom supplier participates in the National Institute of Standards and Technology (NIST) measurement assurance program, which ensures traceability to standard reference materials). The relevant chapters state the phantom's radioactivity concentration at the time the experiments were performed.

Figure 3.3: ^{68}Ge cylindrical phantom

3.2.2 NEMA IEC Body Phantom

The National Electrical Manufacturers Association (NEMA) International Electrotechnical Commission (IEC) body phantom (Figure 3.4) is used in Chapters 4, 9 and 10. This phantom is a fillable torso-sized shell, which optionally contains up to six fillable spheres with varying diameters. This phantom is used to assess PET image quality when performing acceptance tests on newly installed PET systems; image quality must exceed the minimum performance specifications provided by the system manufacturer. This phantom has also been widely used in the literature when assessing PET image quality. Figure 3.4 shows the phantom annotated with the dimensions (spheres are filled with green dye to enhance their visibility). The sphere inner diameters are shown in Table 3.4. An optional 'lung' insert was not used in this study. The background chamber of this phantom was found to contain approximately 10,100ml of water with the spheres in situ.

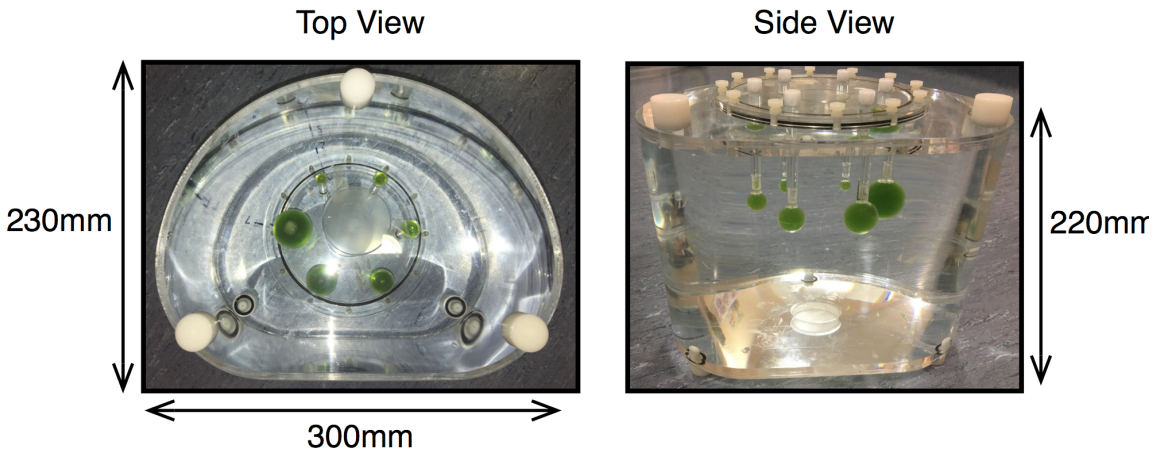


Figure 3.4: NEMA IEC body phantom

Sphere Number	Inner Diameter
1	37mm
2	28mm
3	22mm
4	17mm
5	13mm
6	10mm

Table 3.4: NEMA IEC body phantom sphere diameters

The phantom was used to simulate both 'hot' and 'cold' lesions within background radioactivity. The specific radioactivity concentrations used to fill the phantom are described in each of the relevant chapters.

3.2.3 Spatial Resolution Phantoms

Spatial resolution measurement for iterative reconstruction is not straightforward. Two phantoms were used to evaluate methods of assessing spatial resolution. Both phantoms are assessed and compared in Chapter 7.

3.2.3.1 ^{22}Na Point Source Phantom

A long-lived ^{22}Na spherical point source (0.25mm diameter) embedded in 1cm^3 Lucite was suspended inside the empty shell of the NEMA torso phantom (spheres removed) using a plastic rod. The phantom was then filled with plain water to provide attenuation and scatter material. ^{18}F was also added to the water to provide background activity. To the author's knowledge, no study has attempted to assess spatial resolution by placing a ^{22}Na source inside ^{18}F background activity.

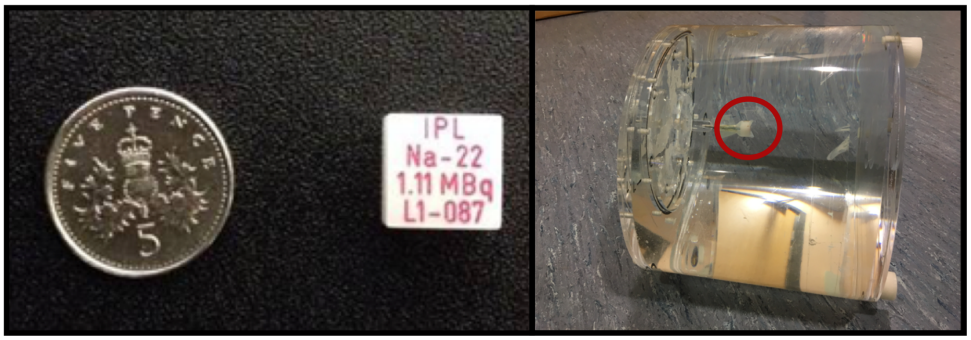


Figure 3.5: ^{22}Na point source inside NEMA IEC body phantom

The following table summarises the differences between ^{22}Na (embedded in Lucite) and ^{18}F (in water).

	^{22}Na (in Lucite)	^{18}F (in water)
Half-Life	2.6 years	109.5 minutes
Branching Factor	90%	97%
Maximum Positron Energy	0.55MeV	0.65MeV
Mean Positron Range in water	Similar to ^{18}F [49]	0.64mm [170]
Calibrated to National Standards	Yes	No

Table 3.5: Comparison between ^{22}Na and ^{18}F

The advantages of placing a ^{22}Na point source inside an ^{18}F phantom are as follows:

- Reproducible point source of known activity, can be used to compare spatial resolution in different parts of the FOV
- Allows spatial resolution to be measured in x, y and z directions using a single acquisition
- Long-lived source placed within a decaying ^{18}F background activity allows different source-to-background measurements to be made without moving the phantom.

However, there are disadvantages to this phantom set-up:

- The point source is separated from the background activity by the cold Lucite casing
- The activity of the point source could not be significantly changed (if required) during the timescale of this thesis.
- The point source and the warm background activity consist of different isotopes. The acquisition system can only use one isotope setting for quantitative corrections. When an ^{18}F background activity was used, the ^{18}F setting was used to correct both the ^{18}F and the ^{22}Na activity, despite the different half-life and branching factors (see Table 3.5).

It was not known if any of these disadvantages would affect the relative differences in spatial resolution as a result of altering the iterative reconstruction parameters prior to experimentation. This is investigated in Chapter 7.

3.2.3.2 Fillable Line Source Phantom

The NEMA 1994 PET image quality phantom contains two line sources and was formerly used for transverse spatial resolution acceptance testing [171]; subsequent versions of NEMA guidelines recommended the use of point sources instead. The stainless-steel line sources have a 1.2mm inner diameter, hold approximately 0.3ml of liquid each, and run the full length of the main phantom chamber (19.6cm diameter cylinder with 20cm length). One line source is positioned at the centre of the phantom and the other is positioned 75mm radially. The NEMA 1994 spatial resolution methodology required that the line source

resolution be measured in air; however, the background chamber of this phantom can be filled with plain water or background activity. The phantom is shown in Figure 3.6 below.

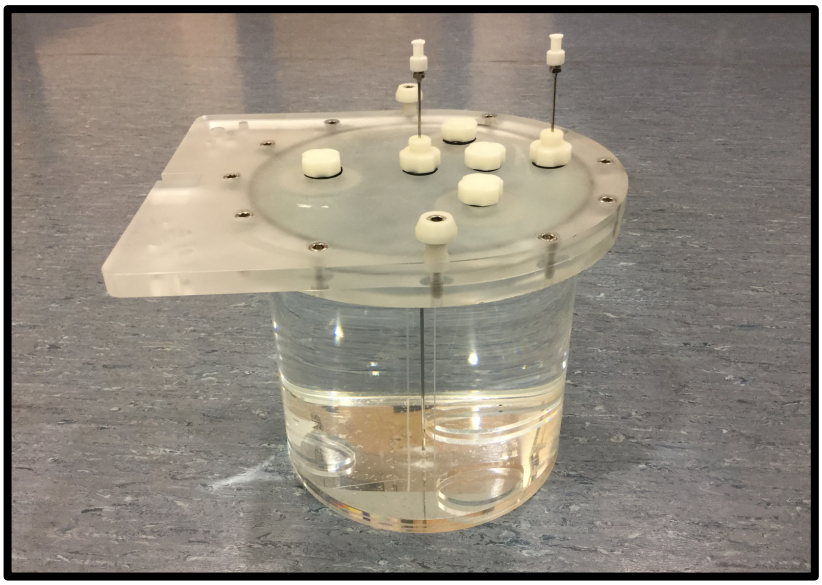


Figure 3.6: NEMA 1994 PET line source phantom

The advantages of this phantom are as follows:

- ^{18}F is used for both the line source and the background activity, so quantitative corrections should be accurate
- The width of the metal line source rods (0.45mm) are less than that of the Lucite surrounding the point source (≈ 5mm), meaning there is less separation between the line source activity and the background activity

However, there are also disadvantages to this phantom set-up:

- The line sources must be refilled for each acquisition, and are therefore not reproducible
- Only one source-to-background ratio can be acquired with each positioning of the phantom (activity can be incrementally added to the background chamber, but the phantom must be moved to achieve this)
- Spatial resolution can only be measured along two of the three image axes using a single acquisition; i.e. in the x and y directions, but not the z direction
- The effect of the stainless-steel rods on spatial resolution is uncertain
- The line source positions are limited to the phantom centre and 75mm offset.

As with the ^{22}Na point source phantom, it was not known if any of these disadvantages would affect the relative differences in spatial resolution as a result of altering the iterative reconstruction parameters prior to experimentation. This is investigated in Chapter 7.

3.3 Image Analysis

Images were analysed primarily using Hermes Medical Systems' commercially available Hybrid Viewer software; however, Matlab was also used to verify some unexpected results in Chapter 6. ROIs were placed on images and used to extract data from specific areas of interest; for example, background radioactivity, simulated lesions in phantom data and areas of interest in patient data. The method and location of ROI placement for each experiment is described in more detail in each of the appropriate chapters.

The use of ROIs in Hybrid Viewer software allowed the extraction of the following statistics:

- **Mean** – the mean voxel value within the ROI
- **Maximum** – the maximum voxel value within the ROI
- **Standard Deviation (SD)** – the standard deviation of all voxels within the ROI
- **SUV_{mean}** – the mean SUV result of all voxels within the ROI
- **SUV_{max}** – The SUV calculated using maximum voxel value within the ROI

3.4 Image Quality Metrics

Extracted ROI statistics were used to calculate the following image quality metrics.

3.4.1 Quantitative Accuracy

As described in Chapter 1, voxel values in reconstructed PET images are presented as radioactivity concentrations (Bq/ml). The ROI Mean and Maximum values therefore represent the average and maximum radioactivity concentrations within the ROI respectively. ROI Mean and Maximum values were used as indications of the quantitative accuracy of PET image data.

3.4.2 Image Noise

Four metrics were used to measure noise within PET images. COV was used to calculate noise when a single large ROI was used. Image Roughness is a variation on COV and was used when multiple ROIs were available. Both COV and Image Roughness measure the voxel-to-voxel variability in the image, i.e. the noise perceived when viewing an individual image [90].

COV of an individual ROI:

$$COV_{ROI}\ (\%) = 100\ \times \left[\frac{SD_{ROI}}{Mean_{ROI}}\right] \qquad \text{Equation 3.2}$$

Image Roughness (IR) for *n* background ROIs is calculated as the average of *n* COVs:

$$IR\ (\%) = \frac{1}{n}\sum_{i=1}^{n} COV_i$$ Equation 3.3 [37]

Background Variation was also measured when multiple ROIs were available. Background Variation measures region-to-region variations throughout the image, which may arise if there are non-uniformities (for example, caused by inaccurate corrections or voxel correlations) [90].

Background Variability (BV) for *n* background ROIs is calculated as the COV of *n* ROI means:

$$BV\ (\%) = 100\ \times \left[\frac{SD_{All\ ROI\ Means}}{Mean_{All\ ROI\ Means}}\right]$$ Equation 3.4 [70]

3.4.3 Lesion Detection

Contrast recovery coefficients (CRCs) were used to measure how well the reconstructed image 'recovers' the expected contrast between the sphere and the background activity, when the actual radioactivity concentrations in both the sphere and background are known. Accurate radioactivity concentrations can only be calculated in well-controlled phantom experiments. Radioactivity concentrations within patients cannot be known exactly, and so CRCs were not used in patient studies. CRCs were calculated for both 'hot' and 'cold' spheres.

Hot Contrast Recovery Coefficients (HCRCs) were calculated using either the sphere ROI's Mean value ($HCRC_{mean}$) or Maximum value ($HCRC_{max}$):

$$HCRC\ (\%) = 100\ \times \frac{\left[\left({C_{hot}}/{C_{bgd}}\right) - 1\right]}{\left[\left({a_{hot}}/{a_{bgd}}\right) - 1\right]}$$ Equation 3.5 [70]

Where:

C_{hot} is the Mean (or Maximum) voxel value in the hot sphere ROI

C_{bgd} is the mean background voxel value

a_{hot} is the true radioactivity concentration in the hot sphere

a_{bgd} is the true radioactivity concentration in the background

Cold Contrast Recovery Coefficient (CCRCs) were calculated using either the lesion/sphere ROI's Mean value ($CCRC_{mean}$) or Maximum value ($CCRC_{max}$):

$$CCRC\ (\%) = 100\ \times \left[1 - \left(\frac{C_{cold}}{C_{bgd}}\right)\right]$$ Equation 3.6 [70]

Where:
C_{cold} is the Mean (or Maximum) voxel value in the cold lesion/sphere ROI
C_{bgd} is the mean background voxel value

SUVs, described in Chapter 1, do not require knowledge of the exact radioactivity concentrations within an ROI and can therefore be used in both phantoms and patients. They do, however, require an accurate record of the activity injected into the patient (minus any post-injection residual) and the patient's body weight (SUVs in the West of Scotland PET Centre are normalised by body weight). Both GEMS and Hermes Medical Systems analysis software automatically calculate SUV_{mean} and SUV_{max} when used to draw ROIs.

$$SUV = \frac{Measured\ Activity\ Concentration\ in\ ROI\ (kBq/ml)}{Injected\ Activity\ (kBq) \Big/ Body\ Weight\ (g)}$$ Equation 3.7

Where:
Measured Activity Concentration in ROI is based on either Mean or Maximum voxel values (for SUV_{mean} or SUV_{max} respectively). Activity concentrations are decay-corrected back to the time of injection.
Injected activity is the total amount of radioactivity injected into the patient (or phantom)
Body Weight is the patient/phantom weight in grams

Finally, the SNR of lesions/spheres were calculated using either the lesion/sphere ROI's Mean (SNR_{mean}) or Maximum (SNR_{max}) value:

$$SNR = 100\ \times \left[\frac{\left(c_{hot} - c_{bgd}\right)}{SD_{bgd}}\right]$$ Equation 3.8 [134]

Where:
C_{hot} is the Mean (or Maximum) voxel value in the hot lesion/sphere ROI
C_{bgd} is the mean background voxel value
SD_{bgd} is the Standard Deviation of the background ROI; when more than one background ROI is used, the average SD of the background ROIs is used.

3.5 Statistics

All graphical and statistical analyses in this thesis were performed using RStudio [172] (version 1.0.143). This graphical user interface was used to run 'R' version 3.3.3. Specific statistical analysis techniques used for each experiment are described in more detail in each of the appropriate chapters.

Chapter 4 : Slice Overlap

Although the primary focus of this thesis is image reconstruction, there were acquisition issues that merited attention; GEMS recommendations were not backed up by evidence. The acquisition issues are therefore addressed before the reconstruction questions are investigated.

A significant aspect of PET data acquisition is the use of overlapping image beds. GEMS recommend a 23% overlap for the Discovery 690 PET-CT system, corresponding to 11 transaxial image slices, for whole-body imaging; however, this recommendation was not supported by the literature. The work in this chapter was therefore undertaken to evaluate the GEMS overlap recommendation by investigating the effect of slice overlap on PET image quality, and to determine an appropriate approach for data acquisition for the remainder of this thesis.

4.1 Introduction

The sensitivity of a PET system, expressed as counts per second per MBq (cps/MBq), represents its ability to detect annihilation radiation. The z-axis sensitivity profile of a scanner operating in 3D mode is triangularly shaped and peaks in the FOV centre. Sensitivity is therefore lower at the edges of the acquired bed position. Whole-body PET scans are typically acquired with overlapping bed positions to compensate for the sensitivity reduction at FOV edges. Data in overlapped regions can then be combined to improve the statistics [173]; however, this increases the required scan time [17]. Overlaps are often quoted as a number of slices or as a percentage of the total bed length. For example, the

GEMS Discovery 690 PET-CT system has 47 slices per bed position. A 23-slice overlap would therefore represent a 49% overlap.

Figure 4.1 shows a simplified schematic of how the z-axis sensitivity profiles of a 6-bed acquisition on the Discovery 690 PET-CT system are altered when different slice overlaps are employed. When no overlap is employed (Figure 4.1(A)), the result is a saw-tooth sensitivity profile with areas of very low sensitivity at the edges of bed positions (approximately 5% of peak sensitivity). A 23%, or 11-slice, overlap (Figure 4.1(B)) improves the relative sensitivity of the overlap areas to ≈ 48% of the maximum. A 49%, or 23-slice, overlap (Figure 4.1(C)) produces an almost uniform sensitivity profile; sensitivity in the overlap areas are ≈ 99% of the maximum. Figure 4.1 also demonstrates that as the overlap size increases, z-axis coverage reduces. This in turn increases the required acquisition time for a given patient length, as more bed positions are required.

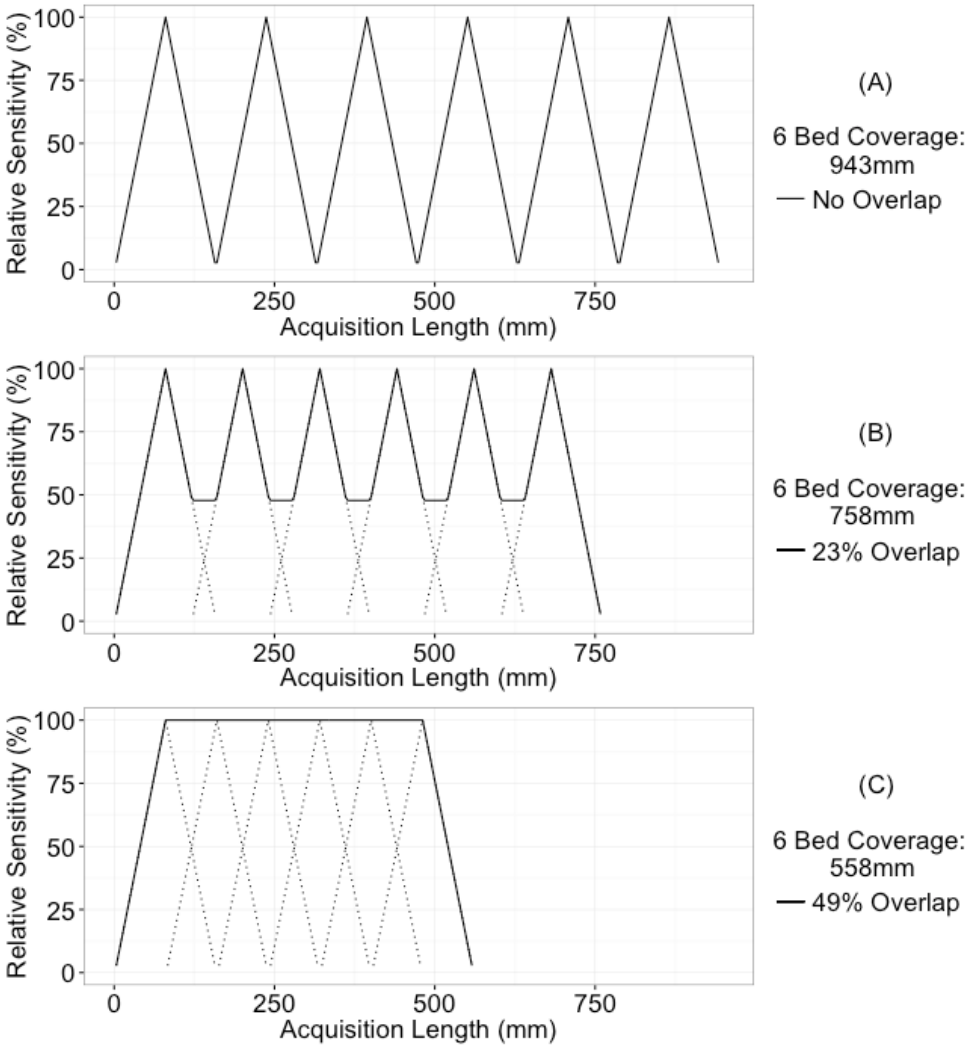


Figure 4.1: Theoretical sensitivity profiles for different overlap settings

Illustration is based upon the GEMS Discovery 690 PET-CT system, which has an axial coverage of 157mm per bed position.

Although the effect of slice overlap on image quality is recognised, the optimal slice overlap has not yet been established. The overlap size that should be employed in 3D PET scanning has been discussed in the literature and the resultant advice varies between studies. A 2005 study by Visvikas et al [174] concluded that altering the size of 3D PET overlaps between 14% and 26% had no effect on image quality. This study, however, had several limitations: the image quality of the 3D patient scans was assessed by comparing them with 2D patient images, no lesion detection gold standard was provided and overlaps larger than 26% were not assessed. A review paper by Humm et al [72] stated an overlap corresponding to between 10% and 20% is necessary for 3D scanning. However, a 2008 study by Tout et al [175] assessed the effects of overlap size upon image noise using a GEMS Discovery STE PET-CT system, and suggested that a minimum overlap of 36% should be used for 3D scanning, with 50% being the optimum overlap size. This study did not assess the effects of overlap upon lesion detection.

The main vendors of clinical PET-CT systems (GEMS, Philips and Siemens) have different approaches to slice overlap. The Siemens Biograph mCT has a fixed overlap of 45% whilst the Philips Gemini has a fixed overlap of 50%. GEMS systems allow operator control of slice overlap; however, the ideal overlap has not been established for GEMS systems. Different overlap sizes are applied among GEMS users, e.g. the Discovery 690 has been used with 15% overlap for whole-body scans [176], 32% overlap for quality control studies [19] and 49% overlap for head and neck scans [176]. It should be noted that the Siemens Biograph mCT Flow introduced the option for continuous bed motion as an alternative to traditional overlapping step-and-shoot bed positions; however, this PET acquisition method is not available to the GEMS Discovery 690 and is therefore not included in this thesis.

Furthermore, overlap size can influence patient injected activities. EANM guidelines for FDG PET-CT imaging [14] can be used to calculate the minimum required radioactivity for patient injections by considering the patient's weight, frame acquisition time and size of overlap employed. A 30% overlap 'cut-off' value is recommended by these guidelines when calculating patient activities: patients being scanned on systems using ≤ 30% overlap would receive double the activity of patients being scanned on systems using > 30% overlap (where patient weight and frame acquisition time are equal). The issues associated with the use of an arbitrary and discrete overlap cut-off are addressed in this chapter.

EANM guidelines also refer to EARL procedures for assessing PET-CT system specific patient FDG activities with respect to image quality [121]. These state that image noise, as measured by the COV in a volume of uniform activity, should not exceed 15%, and that lesion recovery coefficients should remain unbiased as a result of radioactivity reduction.

4.2 Aims

To the author's knowledge, no studies have performed an assessment of the effects of overlap size on image quality on a GEMS Discovery 690 PET-CT system. Furthermore, no studies have assessed the effects of slice overlap on voxel accuracy or lesion detection using phantom studies with known activity concentrations. This chapter therefore aims to assess how the following image parameters are affected by overlap size when the Discovery 690 PET-CT system is used:

1. Image noise (Section 4.4.1)
2. Quantitative accuracy of image activity concentrations (Section 4.4.2)
3. Contrast Recovery Coefficients for hot and cold lesions of varying diameters and varying lesion-to-background ratios (Section 4.5.1)
4. Signal to Noise ratios for low contrast hot lesions of varying diameters (Section 4.5.2)

In particular, this chapter aims to determine if the overlap size recommended by GEMS for the Discovery 690 is sufficient in terms of image quality, by comparing the results obtained using a 23% overlap with those obtained using the maximum possible overlap (49%). Results are discussed with reference to EANM and EARL guidelines.

Finally, an amendment to the EANM guidelines for calculating minimum injected activities is also proposed, based on the findings presented in this chapter.

4.3 Materials and Methods

All acquisitions were performed on the GEMS Discovery 690 PET-CT system. A detailed description of this system is included in Chapter 3.

4.3.1 Phantoms

Two phantom studies were performed.

4.3.1.1 Experiment 1: Uniform ^{68}Ge Phantom

The ^{68}Ge cylindrical phantom described in Section 3.2.1 was used in the first experiment. The phantom's activity was 20.2MBq (3,313Bq/ml in the active volume) at the time of scanning. The phantom was placed in the centre of the FOV. One single-bed PET acquisition was obtained, followed by 23 two-bed PET acquisitions with overlaps varying from one slice (2% overlap) to 23 slices (49% overlap), in one-slice increments. The centre of the overlap was positioned in the middle of the phantom.

4.3.1.2 Experiment 2: ^{18}F Body Phantom with Fillable Spheres

The second phantom study was performed using the NEMA IEC body phantom, which is described in Section 3.2.2. A 4:1 sphere-to-background ratio was used as this is suggested by the NEMA guidelines for assessing image quality [177]. A 2:1 sphere-to-background ratio was also used to simulate lesions of reduced intensity relative to surrounding tissue. Finally, cold lesions were simulated by filling the spheres with plain water. The following acquisitions were performed for each version of the phantom:

- Single-bed acquisition, spheres in FOV centre
- Two-bed, 23% (11-slice) overlap acquisition, spheres in overlap centre
- Two-bed, 49% (23-slice) overlap acquisition, spheres in overlap centre

This sequence of three acquisitions was repeated six times for each phantom as the phantom activity decayed (each phantom was scanned 18 times). Table 4.1 details the measured phantom activity concentrations at the time of the first and last acquisitions (all three phantoms decayed to ≈45% of their original activity during the experiment), as well as the actual sphere-to-background ratios achieved. Initial background activity concentrations were chosen to be similar to that typically found within livers of patients undergoing 400MBq ^{18}F-FDG imaging.

	Background Activity Concentration (Bq/ml)		Sphere Activity Concentration (Bq/ml)		Sphere-to-Background Ratio
	Start	End	Start	End	
Hot Sphere Phantom (4:1)	7,864	3,526	30,622	13,733	3.89:1
Hot Sphere Phantom (2:1)	7,500	3,449	14,899	6,852	1.99:1
Cold Sphere Phantom	9,452	4,133	0	0	N/A

Table 4.1: Body phantom activity concentrations and sphere-to-background ratios

4.3.2 Image Acquisition and Reconstruction Protocol

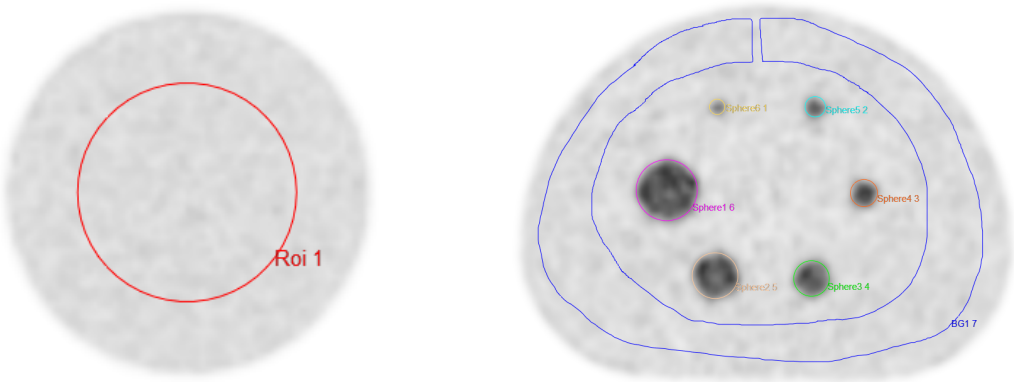
Four-minute acquisition frames were used for all phantom experiments. Images were reconstructed using the parameters suggested by GEMS: TOF and PSF were both enabled, 18 subsets and 3 iterations, 4mm post-reconstruction filter and Standard z-axis filter. This ensured the only differences between reconstructed image sets was the number of overlap slices (and hence sensitivity profiles) and, in the case of the ^{18}F phantom, the total activity in the phantom.

4.3.3 Image Analysis

All ROI analysis was performed using Hermes Medical Systems' Hybrid Viewer software.

4.3.3.1 Experiment 1: Uniform ^{68}Ge Phantom

The single-bed, 47-slice PET image was used to obtain a reference noise value. The central transaxial slice (Slice 24) corresponded to the centre of the axial FOV, where sensitivity is at its maximum, and was therefore used as reference. A 12cm diameter ROI was drawn on this slice and used to measure the mean voxel value, the maximum voxel value and the standard deviation. The mean and maximum voxel values represent activity concentrations in Bq/ml. The mean voxel value and standard deviation were used to calculate the COV of the slice, using Equation 3.2. The 12cm ROI was transferred exactly to the central slice of all 23 two-bed acquisitions with varying overlaps (i.e. in the centre of the overlap area) and used to record/calculate the same data. Figure 4.2 (a) shows the ROI placement.



(a) **(b)**

Figure 4.2: ROI positioning for phantom analyses

(a) 12cm ROI placed on central transaxial slice of ^{68}Ge uniform phantom. (b) Circular sphere ROIs and annular background ROI placed on transaxial slice of body phantom showing centre of spheres.

4.3.3.2 Experiment 2: ^{18}F Body Phantom with Fillable Spheres

The slice containing the centre of the spheres was selected. This slice was at the axial FOV centre for the single-bed scan, and at the overlap centre for the two-bed scans. Data were analysed by drawing ROIs over the largest (37mm diameter) and smallest (10mm diameter) lesion spheres, matching the maximum visible inner diameters on the CT. A large annular ROI was also drawn on this slice to measure the background: this ROI was copied to two adjacent slices either side of the central slice to provide an average background measurement. Mean and maximum voxel values were recorded for the spheres and mean voxel values were recorded for the background. Figure 4.2 (b) shows the ROI placement.

HCRC and CCRC were calculated for all spheres using Equations 3.5 and 3.6 respectively. SNR was also calculated for selected spheres using Equation 3.8.

4.4 Results: Experiment 1

4.4.1 Noise in Uniform ^{68}Ge Phantom

Figure 4.3 shows the central coronal and transaxial slices of a selection of the ^{68}Ge phantom scans. COVs of transaxial slices are also presented.

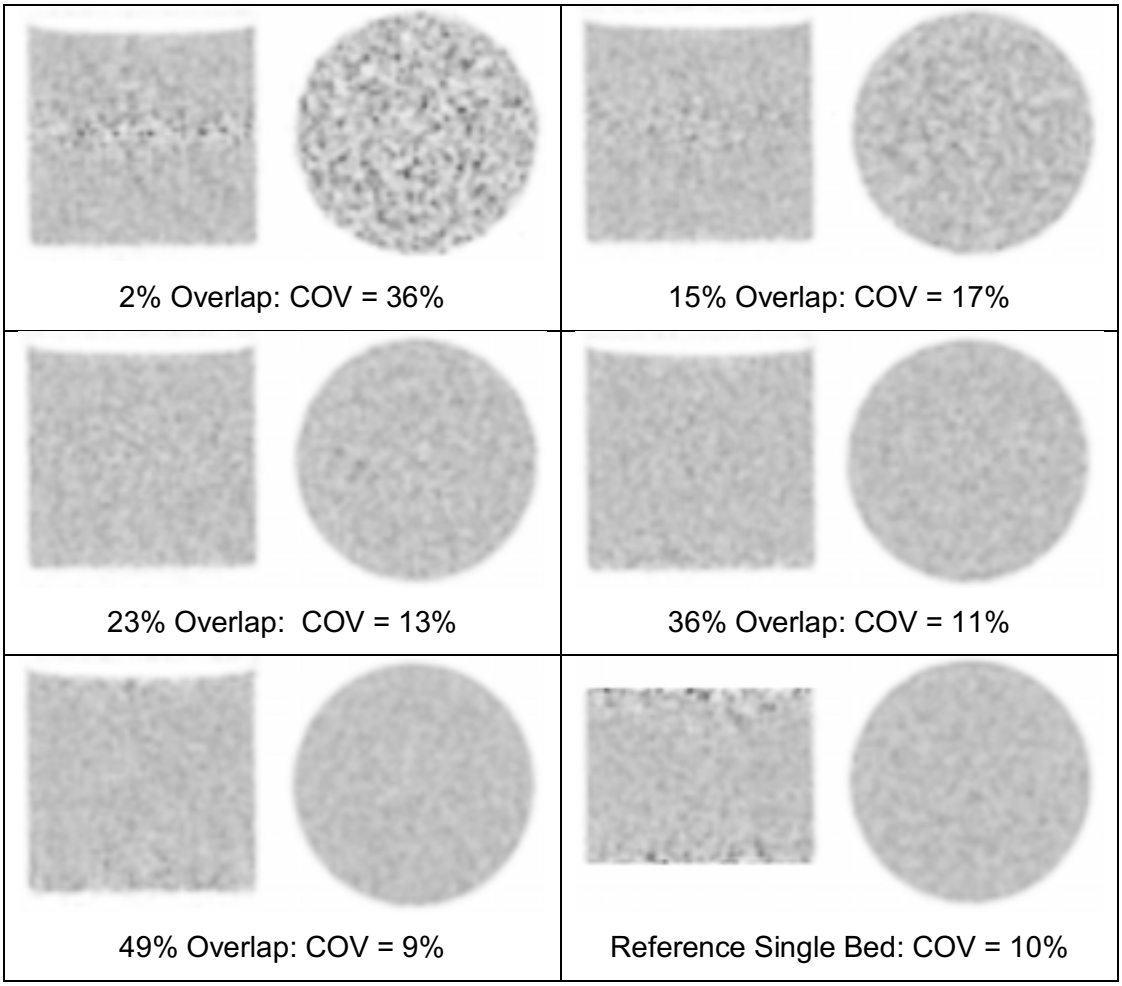


Figure 4.3: Central image slices for ^{68}Ge phantom using different overlaps Both coronal and transaxial slices are shown for each overlap setting.

Figure 4.3 demonstrates the overlap region is clearly visible on the coronal slice when a 2% overlap was used. The overlap is also visible on the 15% overlap coronal slice, but it is not on the 23% overlap. The 2% overlap's transaxial slice is the noisiest both qualitatively and quantitatively (COV = 36%). As the overlap size increases, both coronal and transaxial slices become visibly more homogeneous, with decreasing transaxial COVs.

The 49% overlap central transaxial slice produces similar results to that of the reference single-bed acquisition. The 23% overlap produces COV = 13%, which is 3% greater than the single-bed COV. Applying 23% overlaps would therefore result in a z-axis noise profile

that cycles between a maximum ≈13% and a minimum ≈10% COV. Applying 49% overlaps would result in a consistent noise profile of ≈10% COV.

Qualitative assessment suggests that the 23% overlap has a slightly less homogeneous central transaxial slice than the single-bed or 49% overlap, but it is difficult to visualise the overlap region on the coronal images. This suggests the 23% overlap may be sufficient in terms of image noise, and any noise inconsistencies across the FOV are unlikely to be visualised. The 23% overlap also produces a COV below EARL's 15% limit when four-minute frames are used at these activity concentration levels.

Figure 4.4 plots central transaxial slice COV against overlap size. Single-bed COV (10%) is presented as a local reference and the upper COV limit suggested by EARL (15%) is also included. Finally, the EANM guidelines 30% slice overlap threshold is also indicated.

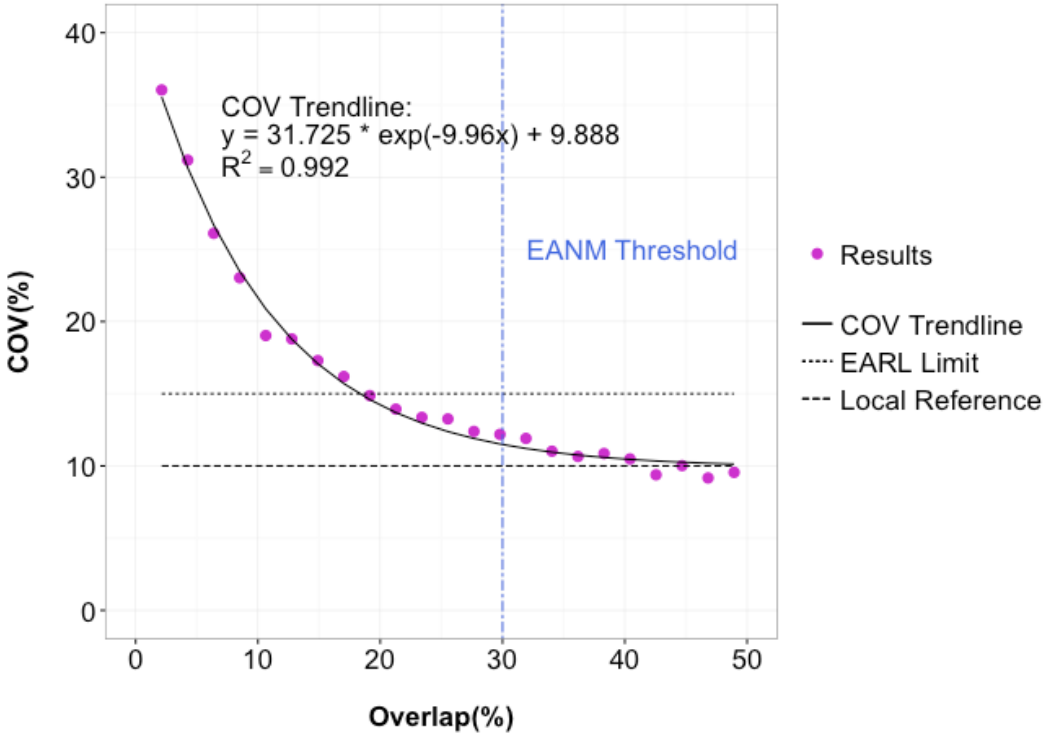


Figure 4.4: Central transaxial slice COV versus overlap size

Local reference COV (10%) and EARL recommended upper limit (15%) are included for comparison. COV Trendline, with associated R^2 value, and EANM slice overlap threshold are also shown.

As overlap size increases the COV initially falls sharply, before a more gradual decrease is observed. There is a clear correlation between central transaxial slice COV and overlap size (offset exponential trendline R^2 = 0.992). Noise exceeds the EARL limit when an overlap of 17% or less is used. The EANM 30% overlap threshold does not appear to correlate with a noticeable watershed in COV performance with respect to overlap size.

The count density of the phantom was then compared to that of patient liver images to determine if these findings were valid for clinical acquisitions. A random group of ten patients undergoing 400MBq ^{18}F-FDG imaging (scanned using the same 4-minute frame times as the phantom study) was found to have an average liver concentration of 9,222Bq/ml (minimum was 6,433Bq/ml). The activity concentration in the phantom (3,313Bq/ml) therefore represents an administered clinical activity of approximately 200MBq, and as such, provides a conservative image noise estimate of not only the West of Scotland PET Centre's routine 400MBq practice, but most other weight-based FDG administration schemes. The 3% COV difference between the 23% and 49% overlaps when four-minute frames are used is therefore unlikely to be clinically significant. Furthermore, the GEMS recommended 23% overlap complies with EARL's 15% COV limit.

4.4.2 Quantification Accuracy of Uniform ^{68}Ge Phantom

Figure 4.5 plots mean and maximum activity concentration measurements against overlap size. Each plot includes the corresponding reference result from the single-bed central transaxial slice (mean 3,328Bq/ml; maximum 4,594Bq/ml). This mean activity concentration measurement was within 0.5% of the known phantom activity concentration at the time of scanning.

Altering overlap size leads to limited variations in mean activity concentrations, which do not appear to correlate with overlap size. The linear trend line demonstrates a poor fit (R^2 = 0.003) and is not statistically significantly different to a zero gradient ($p = 0.791$ from linear regression significance test). There is no positive or negative bias with respect to the reference value (runs test $p = 0.285$).

One may expect the pattern of the maximum voxel values to parallel that of image noise. Figure 4.5 shows that the maximum activity concentrations display a very similar pattern and similar dependence on overlap size as the COV in Figure 4.3, with a similar trendline (R^2 = 0.926). As overlap size increases, the maximum voxel value approaches the reference value.

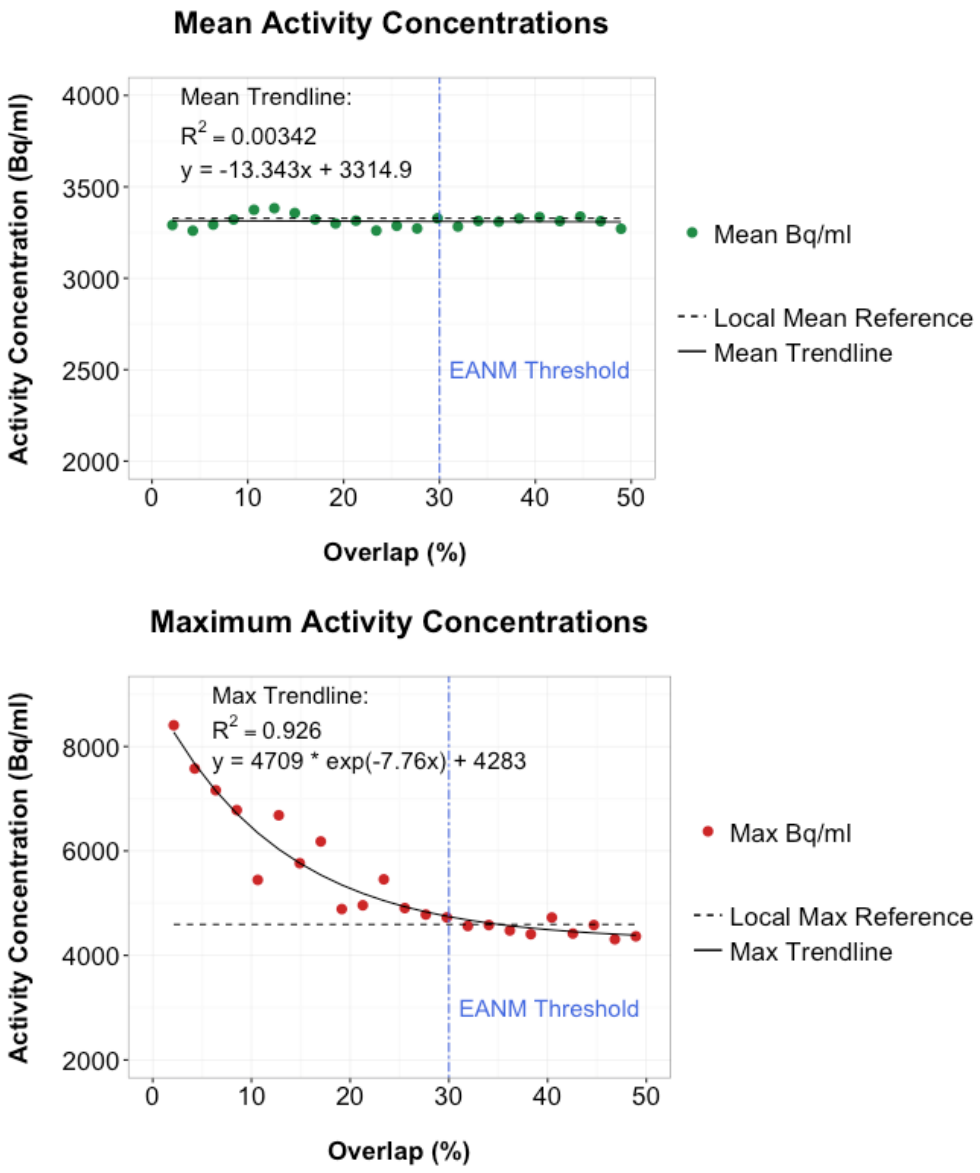


Figure 4.5: Central slice activity concentrations versus overlap size

Dashed lines show local reference values. Trend lines, with associated R^2 values, and EANM slice overlap threshold are also shown.

4.5 Results: Experiment 2

4.5.1 Contrast Recovery for Hot and Cold Spheres

Figure 4.6 shows example hot phantom (≈ 4:1 sphere-to-background ratio) and cold phantom images. Single-bed reference images (spheres positioned at axial FOV centre) are compared with 23% and 49% overlap images (sphere positioned in overlap centre).

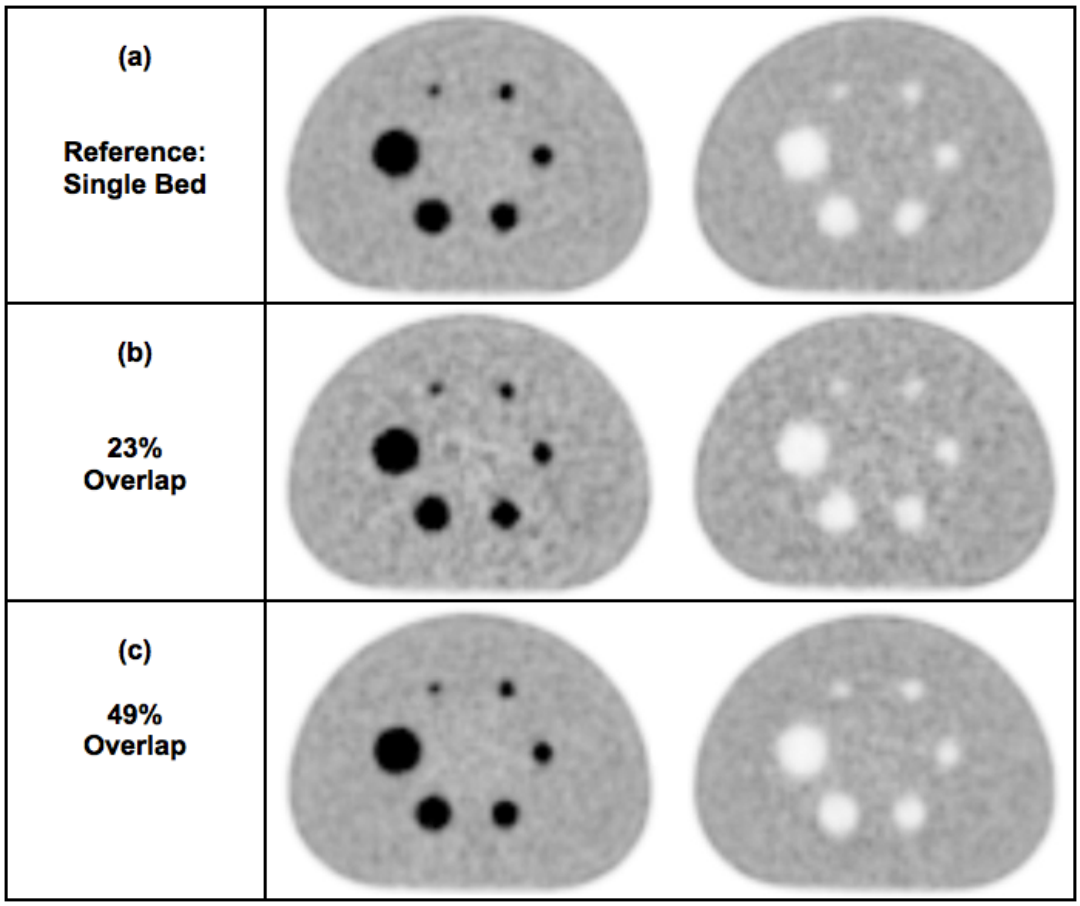


Figure 4.6: Hot and cold sphere phantom images

Images acquired using (a) single-bed acquisition with spheres in FOV centre, (b) two-bed, 23% slice overlap acquisition with spheres in overlap centre, and (c) two-bed, 49% slice overlap acquisition with spheres in overlap centre. Hot phantom has a ≈ 4:1 sphere-to-background activity ratio.

Qualitative assessment of these phantom images indicates that background activity at the overlap centre of the 23% overlap images appears slightly less homogenous than that of the single-bed and 49% overlap acquisitions. However, all six hot and cold spheres are clearly visible for all three overlap settings. This remained the case for all 18 hot phantom and 18 cold phantom acquisitions as the phantom activity decayed.

Average HCRC and CCRC results were calculated for the six spheres for each overlap setting. Table 4.2 compares the average mean and maximum CRCs for the hot and cold spheres for all three overlap settings. Standard Deviations are shown in brackets. Average background region COVs are also presented. As the results are small in number and in multiple, unmatched groups, a Kruskal-Wallis test was applied to determine if there were any statistically significant differences between the three overlap data sets. The resultant p-values are also included in Table 4.2. Finally, the magnitude of the difference in average SUV values is presented to provide some clinical perspective. SUV_{mean} was used for the Mean CRC results, and SUV_{max} was used for the Maximum CRC results. The ΔSUV values

presented in Table 4.2 are the greatest of the three ΔSUV's for each set of results (i.e. the difference between 23% Overlap and 49% Overlap is shown if its magnitude is greater than the difference between either of those results and the Single Bed result).

Hot Spheres (≈ 4:1 Ratio)

	Large Mean HCRC	Large Max HCRC	Small Mean HCRC	Small Max HCRC	Background COV
Single Bed	85.2% (0.016)	110.8% (0.060)	43.0% (0.058)	65.2% (0.090)	8.0%
23% Overlap	87.0% (0.011)	118.8% (0.055)	39.0% (0.037)	60.8% (0.106)	10.3%
49% Overlap	84.8% (0.008)	105.5% (0.046)	39.3% (0.036)	57.4% (0.048)	7.4%
K-W p-value	**0.014**	**0.008**	0.421	0.244	**0.004**
|ΔSUV|	0.06	0.38	0.11	0.22	0.01

Cold Spheres

	Large Mean CCRC	Large Max CCRC	Small Mean CCRC	Small Max CCRC	Background COV
Single Bed	83.4% (0.007)	61.9% (0.017)	46.8% (0.061)	38.6% (0.057)	7.8%
23% Overlap	84.1% (0.003)	63.9% (0.040)	51.8% (0.039)	46.3% (0.044)	9.3%
49% Overlap	83.4% (0.005)	64.5% (0.013)	49.9% (0.036)	43.3% (0.031)	6.9%
K-W p-value	0.080	0.117	0.319	0.061	**0.006**
|ΔSUV|	0.01	0.03	0.05	0.08	0.01

Table 4.2: Contrast recovery coefficients for hot and cold phantoms

Results shown for single-bed, 23% and 49% overlaps. Standard Deviations are included in brackets. Kruskal-Wallis p-values and the difference in SUV_{mean} values are also shown. Background COVs are included for reference.

The Kruskal-Wallis analysis of the contrast recovery results demonstrates only two of the eight sphere result sets had statistically significant differences in contrast recovery between overlap sizes (at the 5% level of significance):

- Large hot sphere's mean HCRC (ΔSUV_{mean} = 0.06)
- Large hot sphere's maximum HCRC (ΔSUV_{mean} = 0.38)

One of these statistically significant results is for the maximum HCRC, which is based on the ROI's maximum voxel value and inherently noisy; it is therefore not surprising that the small increase in noise associated with the smaller overlap has an effect on maximum contrast recovery. The other statistically significant difference was observed in the mean

HCRC of the large sphere: this difference corresponds to a ΔSUV of only 0.06 and is therefore of no clinical significance.

Table 4.2 also includes average background region COV. The 23% overlap has the greatest COV and the 49% overlap has the lowest for both hot and cold phantoms. This is consistent with the uniform phantom COV analysis shown in Figure 4.4. Although the differences in COV for the different overlap settings were statistically significant, the corresponding differences in SUV_{mean} were small (ΔSUV = 0.01 for both hot and cold phantoms). All phantom COVs remained below EARL's 15% limit.

Background activity concentrations for both hot and cold phantoms (Table 4.1) are similar, but generally lower, than those found in patient livers. It is therefore reasonable to consider these results as being applicable to clinical imaging. The maximum observed sphere ΔSUV caused by a change in overlap size was 0.38, whilst the maximum background ΔSUV was 0.01. The effects of slice overlap on CRCs are therefore of no clinical significance for either the ≈ 4:1 sphere-to-background ratio or for the cold spheres.

The ≈ 2:1 sphere-to-background ratio phantom was then analysed. Figure 4.7 shows two example images from each overlap setting. Qualitative assessment suggests the four larger spheres are easily visualised on all acquisitions for all three overlap settings. The smallest sphere is not easily visualised on any overlap setting. Visualisation of the second smallest sphere (13mm diameter), Sphere 5, was inconsistent. When the 49% overlap was used, Sphere 5 was easily visualised on five of the six acquisitions. Sphere 5 was easily visualised on four of the six single-bed acquisitions, and on three of the six 23% overlap acquisitions. It should again be emphasised, however, that these findings only apply to the centre of the overlap region.

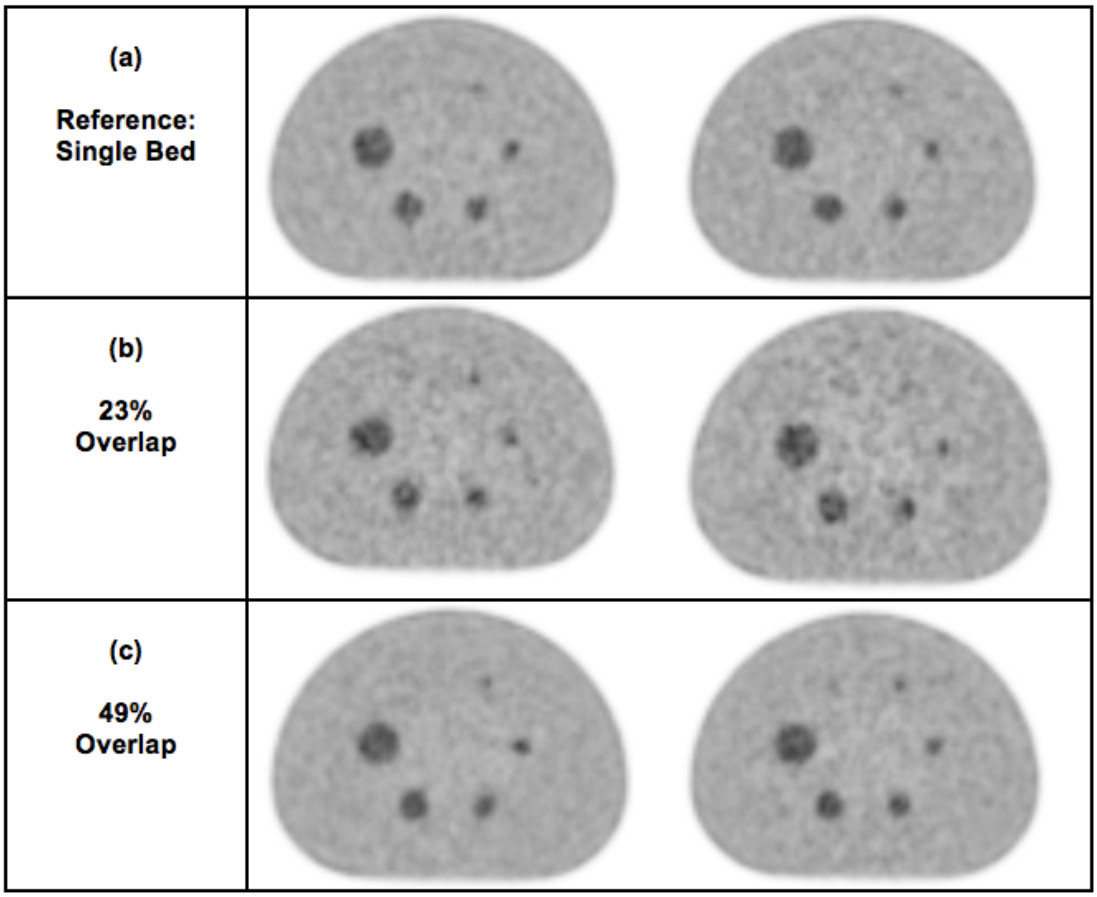


Figure 4.7: Low contrast hot sphere phantom images

Images acquired using (a) single-bed acquisition with spheres in FOV centre, (b) two-bed, 23% slice overlap acquisition with spheres in overlap centre, and (c) two-bed, 49% slice overlap acquisition with spheres in overlap centre. Hot phantom has a ≈ 2:1 sphere-to-background activity ratio.

HCRCs were calculated for Sphere 5 and compared to those of the largest and smallest sphere. Average results are shown in Table 4.3. Standard Deviations are shown in brackets and background COVs are also included. As with the previous data, Kruskal-Wallis p-values and the magnitude of SUV differences are presented.

Mean HCRC

	Large (37mm diameter)	Sphere 5 (13mm diameter)	Small (10mm diameter)	Background COV
Single Bed	77.5% (0.011)	41.4% (0.084)	16.3% (0.094)	8.2%
23% Overlap	76.2% (0.050)	39.1% (0.089)	24.1% (0.138)	10.6%
49% Overlap	78.5% (0.017)	48.5% (0.095)	18.0% (0.056)	7.6%
K-W p-value	0.612	0.236	0.423	**0.007**
|ΔSUV|	0.01	0.08	0.10	0.01

Max HCRC

	Large (37mm diameter)	Sphere 5 (13mm diameter)	Small (10mm diameter)
Single Bed	120.9% (0.090)	58.6% (0.126)	21.9% (0.117)
23% Overlap	130.4% (0.078)	66.7% (0.137)	29.7% (0.130)
49% Overlap	115.9% (0.044)	69.0% (0.132)	25.6% (0.064)
K-W p-value	**0.019**	0.278	0.331
|ΔSUV|	0.16	0.11	0.09

Table 4.3: Contrast recovery coefficients for low contrast hot phantom

Results shown for single-bed, 23% and 49% overlaps. Standard Deviations are included in brackets. Kruskal-Wallis p-values and the difference in SUV_{mean} values are also shown. Background COVs are included for reference.

Kruskal-Wallis analysis showed only one of the six HCRC result sets had a statistically significant difference between overlap sizes (at the 5% level of significance):

- Large hot sphere's maximum HCRC (ΔSUV_{mean} = 0.16)

As stated previously, it is not surprising that the small increase in noise associated with the smaller overlap has an effect on maximum contrast recovery. The difference in SUV_{max} is small, and therefore unlikely to be clinically significant. The background COV results are consistent with both the 4:1 ratio hot phantom and the cold phantom. As before, the phantom background activity concentrations are similar, but generally lower, than those found in patient livers. It is therefore reasonable to consider these results as also being applicable to clinical imaging. The results in Table 4.3 therefore demonstrate that the effects of slice overlap on HCRCs are of no clinical significance for the ≈ 2:1 sphere-to-background ratio.

4.5.2 Signal to Noise Ratios of Low Contrast Spheres

The previous section demonstrated there were no clinically significant differences between contrast recovery coefficients produced using 23% or 49% overlaps. However, qualitative analysis of the low contrast 2:1 ratio phantom suggested that Sphere 5 may be harder to visualise when a 23% overlap is used. Tables 4.2 and 4.3 demonstrated that background noise, as measured by COV, was greatest in the 23% overlap images, as expected; however, image noise also increased as the phantom's activity decayed throughout the course of the experiment. Image noise is known to affect an observer's ability to detect lesions; however, the effects of image noise are not accounted for by contrast recovery coefficients. Signal-to-noise ratios (SNRs), as the name suggests, do account for the effects of image noise and can be considered a more direct measurement of lesion detectability. This section therefore assesses the effects of both slice overlap and phantom decay on SNRs of the low contrast phantom spheres, which are the most challenging to detect.

Figure 4.8 plots SNR of the largest and two smallest spheres in the 2:1 sphere-to-background phantom against phantom activity. Crosses indicate the spheres that could not be visualised (ROIs were positioned based on known sphere locations). Dashed lines show linear regression trend lines, which reflect the relationship between SNR and phantom activity. The gradients of the trend lines, their correlation coefficients and linear regression significance p-values are shown in Table 4.4.

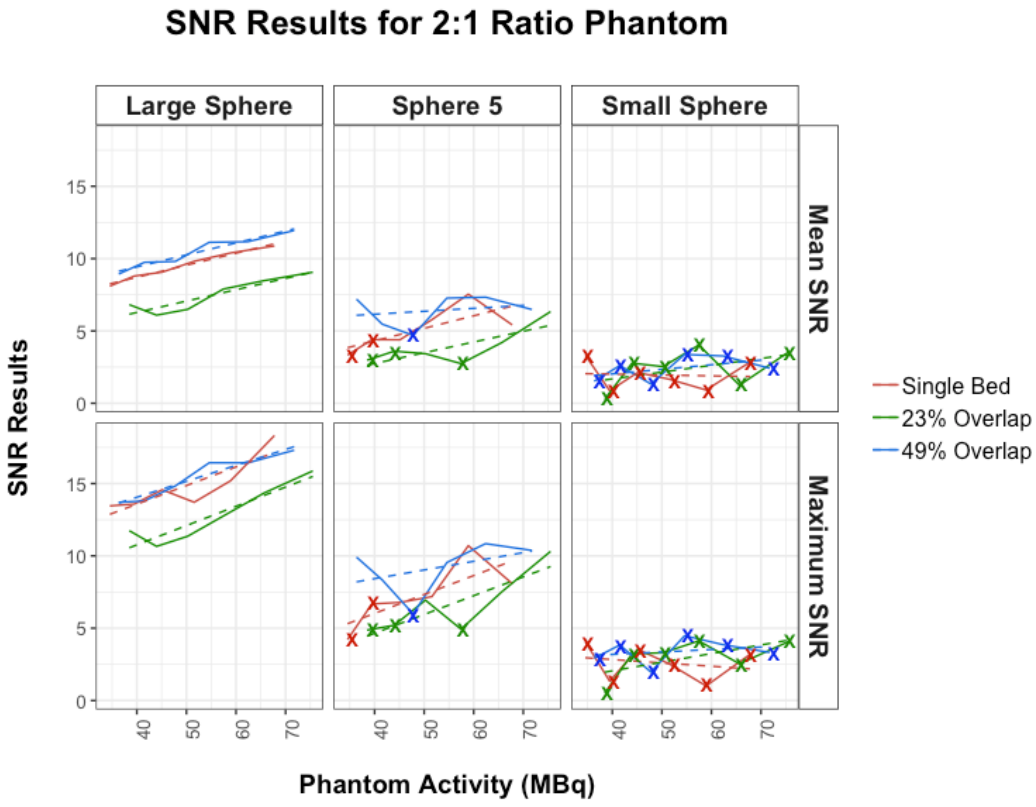


Figure 4.8: SNR results for low contrast spheres

SNR mean and maximum results for largest and two smallest spheres. Phantom acquired with ≈ 2:1 sphere-to-background ratio. Crosses indicate spheres that were difficult to visualise.

One would expect each sphere's SNR to decrease as the phantom activity decreases, due to the associated increase in statistical noise. Figure 4.8 demonstrates the Large Sphere follows this expected pattern, confirmed by the linear regression analysis: all Large Sphere trend lines have reasonable fits (minimum R^2 = 0.746) and gradients that are statistically significantly different to zero (maximum p-value = 0.027). The smallest sphere, however, does not demonstrate any correlation between SNR and phantom activity: no Small Sphere trend lines are significantly different to a zero gradient. Small Sphere measurements are inherently noisier than those of the Large Sphere due to its smaller size; this increased noise, combined with the low sphere contrast ratio, is likely to mask any underlying trend in the data. Sphere 5 appears to largely demonstrate the expected relationship between SNR and phantom activity, as all six data series produce positive linear gradients. However, only one series, 23% Overlap Maximum SNR, produces a gradient that is statistically significantly different to a zero gradient (p-value = 0.033). Sphere 5 measurements are also inherently noisier than those of the Large Sphere, which may explain why the underlying trends are not statistically significant.

	Single Bed			23% Overlap			49% Overlap		
	R^2	Grad	p	R^2	Grad	p	R^2	Grad	p
Large Mean	0.982	0.083	**<0.001**	0.839	0.078	**0.010**	0.936	0.083	**0.002**
Large Max	0.746	0.129	**0.027**	0.864	0.133	**0.007**	0.926	0.110	**0.002**
Sphere 5 Mean	0.544	0.087	0.094	0.619	0.074	0.063	0.060	0.020	0.640
Sphere 5 Max	0.589	0.131	0.075	0.718	0.130	**0.033**	0.201	0.060	0.373
Small Mean	0.005	-0.006	0.890	0.215	0.047	0.354	0.243	0.032	0.320
Small Max	0.055	-0.023	0.655	0.379	0.060	0.194	0.066	0.017	0.623

Table 4.4: Linear regression results for low contrast phantom SNR

Correlation coefficients, gradients and p-values of a significance test for linear regression are shown. Phantom acquired with 2:1 ratio.

With respect to overlap size, both Mean and Maximum SNR results for the Large Sphere in Figure 4.8 demonstrate clear differences between the 23% Overlap and those of the 49% Overlap and Single Bed: the 23% Overlap produces the lowest results while the 49% Overlap largely produces the greatest results. Figure 4.8 also suggests there are differences in Sphere 5's SNR as a result of overlap size; however, the differences are not as apparent as they are for the Large Sphere (23% Overlap results are generally the lowest, while 49% Overlap results are generally the greatest). Slice overlap does not appear to have any obvious effect on the Small Sphere results.

Kruskal-Wallis analysis was applied to the data to determine if overlap size had any statistically significant effect upon SNR. The resultant p-values are shown in Table 4.5, which also includes average SNR results. The magnitude of SUV_{mean} differences are once again presented.

Mean SNR

	Large (37mm diameter)	Sphere 5 (13mm diameter)	Small (10mm diameter)
Single Bed	9.51 (1.04)	5.14 (1.46)	1.96 (1.00)
23% Overlap	7.47 (1.18)	3.88 (1.30)	2.40 (1.40)
49% Overlap	10.45 (1.13)	6.41 (1.10)	2.42 (0.86)
K-W p-value	**0.007**	**0.022**	0.700
|ΔSUV|	0.01	0.08	0.10

Max SNR

	Large (37mm diameter)	Sphere 5 (13mm diameter)	Small (10mm diameter)
Single Bed	14.81 (1.85)	7.27 (2.12)	2.61 (1.23)
23% Overlap	12.79 (1.98)	6.63 (2.12)	2.96 (1.34)
49% Overlap	15.42 (1.52)	9.17 (1.78)	3.40 (0.89)
K-W p-value	0.064	0.128	0.523
|ΔSUV|	0.16	0.11	0.09

Table 4.5: Average SNR results for low contrast phantom

Results are shown for the largest and two smallest spheres acquired using single-bed, 23% and 49% overlaps. Standard Deviations are included in brackets. Kruskal-Wallis p-values and the difference in SUV_{mean} values are also shown. Average background COVs are also included. Phantom acquired with 2:1 ratio.

Kruskal-Wallis analysis demonstrates two of the six SNR result sets have statistically significant differences between overlap sizes (at the 5% level of significance).

- Large Sphere's mean SNR (ΔSUV_{mean} = 0.01)
- Sphere 5's mean SNR (ΔSUV_{mean} = 0.08)

These correspond to very small differences in sphere SUV_{mean}; however, sphere SUV_{mean} values do not reflect the effects of noise.

The 23% overlap has been shown to have the greatest noise, as measured by background COV, whilst the 49% overlap has the lowest noise. One would therefore expect the 23% overlap to have the lowest SNR and the 49% overlap to have the greatest SNR. The results in Table 4.5 largely confirm this theory: the 49% overlap has the greatest SNR for all six result sets, whilst the 23% overlap has the lowest SNR for 4 of the 6 result sets. The smallest sphere results were too noisy to demonstrate any relationship between SNR and slice overlap.

The crosses in Figure 4.8 indicate spheres that were difficult to visualise qualitatively. The largest sphere was visible on every acquisition, while the smallest sphere wasn't visualised on any of the acquisitions. The size of the overlap therefore did not affect the visualisation of the Large or Small Sphere.

Sphere 5 was not visualised on six of the eighteen acquisitions. The 23% Overlap accounted for three of the six non-visualised spheres, while Single Bed accounted for two and 49% Overlap accounted for only one. There was a statistically significant difference in overlap performance for the mean SNRs. However, it should be noted that Sphere 5's detectability was only compromised once the phantom background activity had decayed to approximately 5,700Bq/ml (lower than typically found in patient livers). As with the Small Sphere, the results appear to suggest that to visualise a small lesion (13mm diameter), the mean SNR greater must exceed approximately 5 (if not greater).

4.6 Discussion

Figures 4.3 and 4.4 compared the single-bed reference noise level (COV ≈ 10%) to noise levels achieved when overlap size varies. These figures demonstrate that as overlap size increased, noise in the overlap region decreased. Noise in the 49% overlap (COV ≈ 10%) was similar to the single-bed reference, while the 23% overlap (COV ≈ 13%) had greater noise. This illustrates the theoretical advantage of using larger overlaps. However, it was difficult to visualise the increase in noise when assessing the images qualitatively. This suggests that the 23% overlap is sufficient in terms of image noise when this particular combination of acquisition time, reconstruction parameters and activity concentration is used, as any noise inconsistencies across the z-axis FOV are unlikely to be visualised. Additionally, the 23% overlap COV did not exceed the 15% COV limit recommended by EARL.

A comparable study by Tout et al [175] used a similar ^{68}Ge phantom (activity concentration approximately 5,161Bq/ml) to assess COV on the GEMS Discovery STE scanner (an older PET-CT model that used BGO crystals and had no TOF or PSF capabilities). Coronal phantom images demonstrated increased noise in the overlap region for overlaps of 32% or less, which led the authors to conclude a minimum 36% overlap should be used (with 50% overlap being the optimal choice). However, the authors also stated slice overlap should be matched to the capabilities of each system in question. The work undertaken in this chapter demonstrated that the overlap region was not visible on coronal phantom images when a 23% overlap was used on the Discovery 690, and this was assessed using phantom images with a ≈ 35% lower count density than the Discovery STE study.

Standardised Uptake Values (SUVs) are used in clinical PET to characterise lesion uptake and are based upon voxel values. Figure 4.5 illustrates the effect overlap size has on activity concentration accuracy for both mean and maximum voxel values. As mean voxel values were not affected by overlap size, a smaller overlap can be used without penalty with respect to mean voxel values or SUV_{mean} measurements.

However, maximum voxel values, and therefore SUV_{max}, in the overlap area are affected by overlap size, as they are directly affected by image noise. The 49% overlap maximum voxel value was similar to that of the single-bed reference, while the 23% overlap maximum voxel value was 19% greater than that of the single-bed reference (for the particular combination of phantom activity and acquisition time used). Maximum voxel values and SUV_{max} measurements should always be used with caution regardless of overlap size, as this type of measurement is directly affected by image noise [178].

Contrast Recovery Coefficients (CRCs) for different sized spheres and different sphere-to-background ratios were also analysed. EARL guidelines state that lesion recovery coefficients should remain unbiased. Overlap size was found to have no clinically significant effect on any of the CRCs. However, as with SUV_{max} measurements, maximum CRCs should be used with caution.

CRCs alone are not a reliable measure of lesion detectability, as demonstrated by the low contrast 2:1 sphere-to-background phantom results. The HCRC calculations suggested there was no clinically significant difference between the 23% and 49% overlap results; however, qualitative analysis demonstrated a difference between 23% and 49% overlaps when visualising Sphere 5. This was confirmed by the SNR calculations. It is important to note that Sphere 5's visibility was only compromised on the 23% overlap images when the phantom's background activity concentration had decayed below 5,700Bq/ml. The resultant noise increase made the sphere difficult to visualise.

Phantom experiments in this chapter were performed using four-minute frames and were compared with activity concentrations within patients injected with 400MBq ^{18}F-FDG. Reducing frame times or injected activities may further compromise lesion detectability. However, the count density of the ^{68}Ge phantom was shown to be representative of patients receiving 200MBq injections; the experiment therefore provided a conservative image noise estimate for a wider range of FDG administration schemes. Furthermore, the visualisation of the low contrast Sphere 5 was only compromised when the phantom activity concentration fell below 5,700Bq/ml. This is 11% lower than the lowest observed patient liver activity concentration for a 400MBq ^{18}F-FDG scan, and 37% lower than the average patient liver activity concentration. This suggests frame times may be reduced below 4

minutes (or injected activities could be reduced below 400MBq) whilst using a 23% overlap without compromising the visualisation of low contrast lesions; however, further work should be undertaken to justify any protocol changes.

There are practical reasons that favour the use of smaller overlaps to maximise image quality. A typical scan length of 900mm on the GEMS Discovery 690 requires 8 bed positions when a 23% overlap is used. This increases to 11 bed positions if the overlap is increased to 49%. Increasing the number of bed positions, and hence scan duration, has practical implications:

- Patients would have to lie still for a longer period of time. This increases the risk of physical and physiological movement, which can degrade image quality. Severe motion leads to repeated imaging and additional CT exposure, which would increase the total radiation dose burden to the patient.
- The daily patient throughput of the imaging centre would be reduced. Although total imaging time for a particular patient depends on the prescribed imaging extent (and hence the patient's height), each planned patient time-slot would need to be increased to account for this extra imaging time.

A 23% overlap, with four-minute frames, has been employed in the West of Scotland PET Centre since the Discovery 690 system was installed in 2011, as per manufacturer guidance at that time. No overlap-related artefacts have ever been reported. Figure 4.9 shows coronal images of three patients with varying BMI who were all imaged under the same conditions: no overlap regions are visible on any of these images.

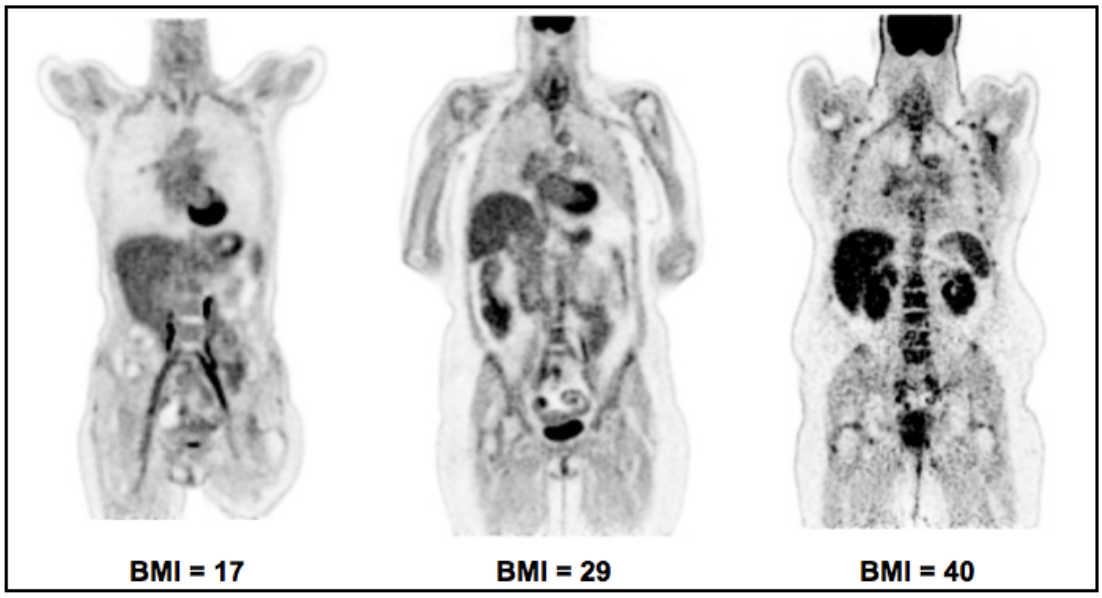


Figure 4.9: Coronal slices of three PET patients with varying BMI

All three patients were injected with 400MBq ^{18}F-FDG and scanned at 60 minutes post-injection. PET acquisitions obtained using four-minute frames with 23% slice overlap.

4.7 Proposed Amendment to European Guidelines

The 2014 EANM guidelines outline specifications for calculating weight-based injected activities, which take frame time and overlap size into consideration. Both linear and quadratic schemes are outlined in Equations 4.1, 4.2, 4.3 and 4.4.

$$EANM\ Min\ Activity\ (Linear) = 14 * \frac{Patient\ Weight\ (kg)}{Frame\ Time\ (mins)}, \quad Overlap\ \leq 30\% \qquad \text{Equation 4.1}$$

$$= 7 * \frac{Patient\ Weight\ (kg)}{Frame\ Time\ (mins)}, \quad Overlap > 30\% \qquad \text{Equation 4.2}$$

$$EANM\ Min\ Activity\ (Quadratic) = 1{,}050 * \frac{\left(Patient\ Weight\ (kg)/_{75}\right)^2}{Frame\ Time\ (mins)}, \quad Overlap \leq 30\% \qquad \text{Equation 4.3}$$

$$= 525 * \frac{\left(Patient\ Weight\ (kg)/_{75}\right)^2}{Frame\ Time\ (mins)}, \quad Overlap > 30\% \qquad \text{Equation 4.4}$$

According to these guidelines, an average 70kg patient to be imaged using four-minute frames and <30% overlap would require a minimum injected activity of 245MBq using the linear model, or 228MBq using the quadratic model. If the overlap was increased >30%, the minimum injected activity would be halved to 122.5MBq (linear) or 114MBq (quadratic). The

work undertaken in this chapter demonstrated this 30% EANM overlap threshold does not correlate with a noticeable watershed in COV performance. Figure 4.4 also demonstrated the relationship between COV and overlap size can be modelled as an offset exponential relationship.

An alternative calculation scheme is therefore proposed which accounts for the measured relationship between image noise and slice overlap:

$$Minimum\ Activity, Linear\ Model\ = -10\ ln(x)\ *\ \frac{Patient\ Weight\ (kg)}{Frame\ Time\ (mins)} \quad \text{Equation 4.5}$$

$$Minimum\ Activity, Quadratic\ Model \quad \text{Equation 4.6}$$

$$= -750\ ln(x)\ *\ \frac{\left({Patient\ Weight\ (kg)}/{75}\right)^2}{Frame\ Time\ (mins)}$$

where *x* is the percentage overlap size expressed as a decimal (e.g. x = 0.23 for 23% overlap).

Although the relationship between COV and overlap size was modelled using an offset exponential in Figure 4.4, the application of the COV/overlap relationship has been simplified in the above equations for clinical use by using the natural log of the overlap and a single multiplicative constant (the use of an offset exponential with three constant terms may be impractical in the clinical setting). Departments using a single fixed overlap for all scanning can simplify these equations further by pre-calculating the '-10ln(x)' and/or '-750ln(x)' terms and substituting them as replacement constants to the current EANM calculation scheme.

The multiplicative constants in these equations were chosen to produce values within 1% of the EANM models at the maximum possible overlap (50%). This reduces the impact of the proposed guidance change on centres with scanners that employ large fixed overlaps. As the overlap size decreases, the minimum activity increases according to the logarithmic model to account for the associated increase in image noise.

Furthermore, a minimum overlap of 23% for the GEMS Discovery 690 is also proposed in these models for the following reasons:

- The 23% overlap is recommended by GEMS
- This study has demonstrated that a 23% overlap is safe for clinical use

- Detectability of small lesions in low count studies may be compromised when overlaps less than 23% are used, as demonstrated in Section 4.5 (smaller overlaps would further degrade the SNR)
- With respect to the proposed models for calculating minimum injected activities, this overlap restriction also prevents the use of large injected activities to compensate for the increased noise levels associated with smaller overlaps.

Figure 4.10 illustrates the effect these amendments would have on the calculated minimum injected activities for a 70kg patient to be imaged using 4-minute acquisition frames. When overlaps of between 30% and approximately 40% are used under the current guidelines, reduced image quality in overlap areas may compromise lesion detectability.

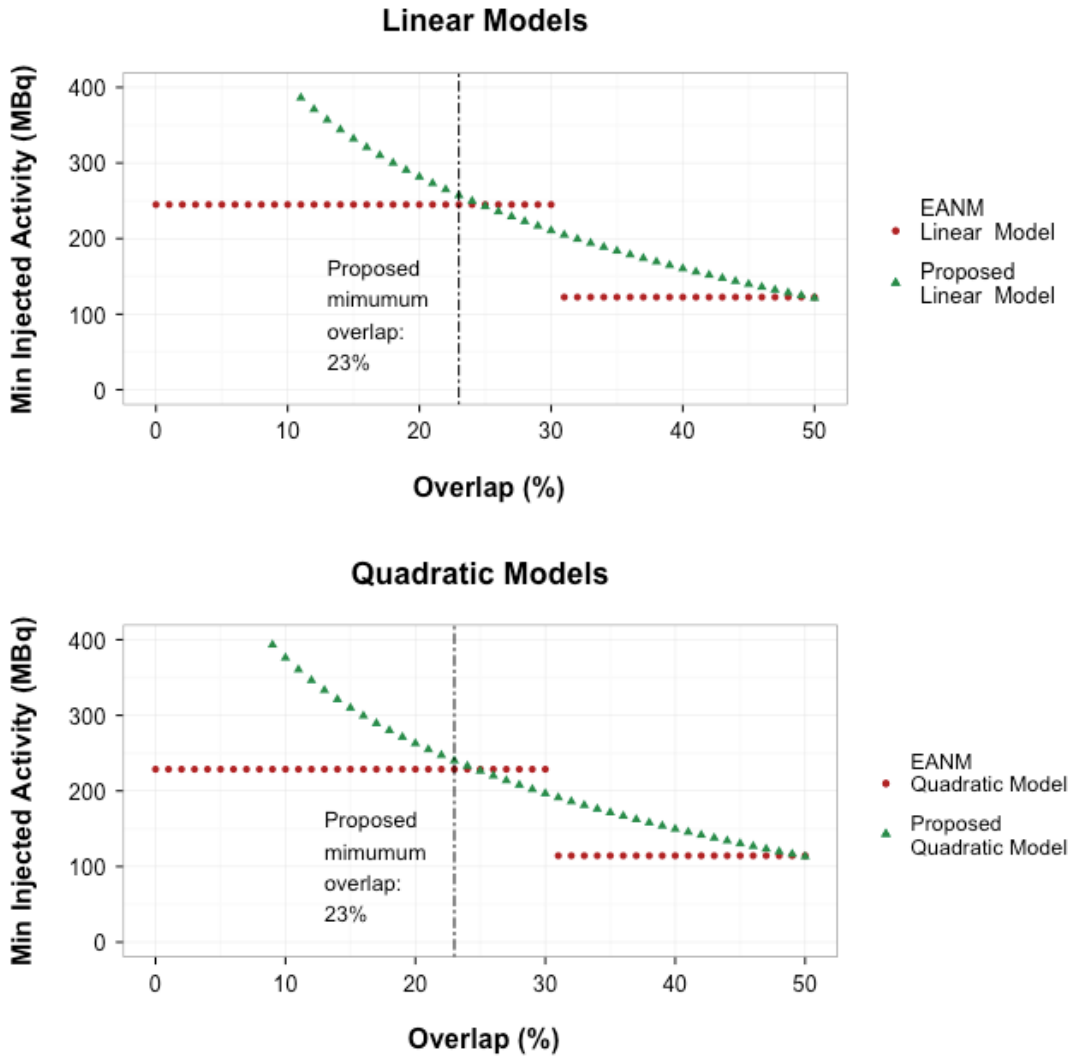


Figure 4.10: Illustration of proposed EANM amendment

Comparison of proposed weight-based injected activity models with current EANM models, for a 70kg patient imaged using 4-minute acquisition frames. Proposed models are restricted to a minimum overlap of 23%.

4.8 Conclusions

A review of the literature revealed a paucity of studies assessing the effects of slice overlap on PET image quality. Furthermore, the GEMS recommended 23% overlap for the Discovery 690 was not supported by any peer reviewed studies. This chapter therefore aimed to characterise the effect of overlap size on image quality, in terms of image noise, voxel accuracy, lesion contrast recovery and detectability. In particular, this chapter aimed to determine if there was evidence to justify the use of a 23% overlap using the Discovery 690 by comparing 23% overlap results with those obtained by the maximum possible overlap (49%).

When 23% overlap images are compared to 49% overlap images:

- There is minimal detectable influence on clinical image quality in terms of noise, mean activity concentrations or mean contrast recovery of lesions
- There is no detectable influence on SUV_{mean} values
- Detectability of small, low contrast lesions may be affected in low count studies, not typically seen in local routine clinical practice, when a 23% overlap is used
    - This is a worst-case scenario, which applies only when lesions are located in the overlap area
- Maximum contrast recovery and SUV_{max} values should be used with caution

The marginal discernible benefits of increasing the slice overlap from 23% to 49% would likely be offset by the disadvantages of increasing clinical scan times. No overlap-related artefacts have been reported to date when 23% overlaps have been employed locally for clinical scans. Therefore, this chapter concludes a 23% overlap may be used on the GEMS Discovery 690 without detriment to clinical imaging, unless the combination of administered activity and frame time is unusually low. All practical work undertaken throughout the remainder of this thesis therefore used an overlap of 23%.

This chapter also proposes an amendment to the EANM guidelines which accounts for the relationship between image noise and slice overlap when calculating minimum patient injection activities.

Some of the work performed in this chapter was published in the Physics in Medicine and Biology Journal in January 2016 [179].

Chapter 5 : Quantitative Accuracy in Reconstructed Images of a Uniform Phantom

Quantitative image metrics used throughout this thesis, and for PET imaging in general, are calculated using reconstructed voxel activity concentrations, the accuracy of which are dependent upon appropriate calibration of the PET system. The GEMS procedure for calibrating the Discovery 690 uses a combination of reconstruction parameters which differ from their suggested clinical reconstruction parameters; the implications of this are not discussed in GEMS documentation. Furthermore, there is little discussion of the importance of reconstruction parameters used in PET calibration procedures in the wider literature. Surprisingly, there is also no consideration of the effects of fundamental changes to reconstruction algorithms associated with TOF and PSF corrections. This chapter therefore investigates the effects of reconstruction parameters upon data used for PET calibration.

5.1 Introduction

Voxel values in reconstructed PET images are the result of a complex reconstruction process with many user-defined parameters, as described in Chapter 1. Voxel values are presented as radioactivity concentrations, as described in Section 1.5.1; Well Counter Calibration (WCC), or absolute activity calibration (AAC), is used to convert image voxel values into activity concentrations [65], [76]. Whilst qualitative analysis of clinical PET images is common practice, quantitative analysis has also been shown to be useful in the evaluation of tumour response to treatment, as discussed in Chapter 2. Regions of interest (ROIs) are often used to produce quantitative analysis of reconstructed images, for both clinical and research purposes; for example, to measure uptake within a lesion or noise within background activity.

Although voxel activity concentrations are not typically quoted directly in clinical PET reports, they are used to calculate clinical metrics such as SUV, TLG and MTV [141], [180], [181]. Both mean and maximum voxel values within ROIs are of clinical relevance, as both can be used to calculate these clinical metrics. For example, as described in Section 1.5.2, maximum voxel values are used to calculate SUV_{max} measurements, while SUV_{mean} and SUV_{peak} use averaging of voxel values in the relevant region.

Accurate WCC is therefore of critical importance to the clinical interpretation of PET images. The American College Radiology Network (ACRIN) Centre of Quantitative Imaging Excellence (CQIE) Manual of Procedures [182] states that WCC should be performed at least every three months, after scanner upgrades, after new set-ups, and after modifications to the radionuclide calibrator. EANM quality control guidelines [125] state that WCC should be performed as per manufacturer's procedures. GEMS guidelines for the Discovery 690 PET-CT system state that WCC should be performed every three months and following any upgrades, as per the ACRIN procedures.

The GEMS procedure for WCC involves filling a 20cm uniform phantom with a known ^{18}F activity concentration of approximately 3,500Bq/ml, performing a WCC phantom acquisition/reconstruction and generating correction factors based on the measured data. Verification of the new correction factors involves performing and analysing a clinical acquisition/reconstruction of the same phantom; mean voxel values should be within ±10% of the expected value (given the known decay-corrected phantom activity concentration). The GEMS WCC protocol uses different reconstruction parameters from those suggested by GEMS for clinical reconstructions, as summarised in Table 5.1.

	GEMS Recommendations	
	WCC	**Clinical**
Effective Iterations	48	54
FOV	250mm	700mm
Voxel Size (x & y axes)	1.95mm	3.65mm
Post-Reconstruction Gaussian filter	6.4mm	4mm
Z-Axis Filter	None	Standard
TOF	No	Yes
PSF	No	Yes

Table 5.1: GEMS reconstruction parameters for WCC and clinical imaging

The rationale for using this particular combination of reconstruction parameters for WCC is not explained in GEMS documentation. GEMS were therefore contacted directly for more

information about the Discovery 690's WCC protocol, and in particular, for an explanation as to why different reconstruction parameters are used for WCC and clinical scans. Dr Charles Stearns, principle engineer at GEMS (personal communication, 21st July 2014) stated that quantitative values produced by the Discovery 690 for large, uniform objects, like the 20cm diameter uniform phantom used for WCC, are insensitive to the following reconstruction parameters, provided the relevant conditions are met:

- Effective iterations (provided 'enough' iterations are performed to obtain 'decent' image quality)
- Filtering (unless at the very edge of the field of view)
- Voxel size

Dr Stearns also stated that the algorithm type (which determines if TOF or PSF are applied) is 'calibrated out' by a correction factor in the WCC reconstruction, which makes the output constant across all algorithm types.

The impact of using different combinations of reconstruction parameters for WCC and clinical acquisitions has not been widely discussed in the scientific literature. A 2013 textbook by Prekeges [35] states that separate WCC correction factors are required for different reconstruction algorithms, which appears to contradict the advice given by GEMS. Other textbooks and publications that explain the basic principles of WCC do not discuss the impact of reconstruction parameters [60], [125], [183]. There appears to be a widespread assumption that using different reconstruction parameters for WCC and clinical scans has no detrimental impact upon quantitative accuracy; however, there is no published evidence explicitly stating this is the case. This is surprising when one considers how critical accurate WCC is to quantitative accuracy of PET images.

5.2 Aims

The first aim of this chapter is to assess the differences in voxel values between reconstructions produced using the WCC parameters and the GEMS suggested clinical parameters, outlined in Table 5.1, for a uniform phantom. This will test the GEMS assertion that the current WCC reconstruction protocol is suitable for use with the GEMS suggested clinical reconstruction parameters.

The second aim of this chapter is to characterise the effects of altering each of the reconstruction parameters in Table 5.1 upon voxel accuracy in a uniform phantom. This will test the wider GEMS assertion regarding the insensitivity of voxel values to effective iterations, filtering and voxel size (within the limits stated by GEMS). The effects of TOF and PSF will also be assessed, for completeness.

5.3 Materials and Methods

All acquisitions and reconstructions were performed on the Discovery 690 PET-CT system. A detailed description of this system is included in Chapter 3.

5.3.1 Phantom

Although a uniform ^{18}F phantom is used for performing WCC on the Discovery 690, this chapter used the uniform ^{68}Ge phantom described in Section 3.2.1 for two main reasons:

- This chapter assesses voxel accuracy of reconstructed images in lieu of performing WCC calibrations with different reconstructions. The ^{68}Ge phantom is calibrated to NIST standards and so provides an accurate standard to within ±3% tolerance. An ^{18}F phantom would introduce measurement errors of up to ±5% associated with the radionuclide calibrator [184] with both the pre-injection and post-injection syringe measurements.
- The ^{68}Ge phantom is known to have a uniform activity distribution with no air bubbles. The ^{18}F water phantom is subject to potential air bubbles and non-uniform mixing of the activity throughout the phantom volume.

The characteristics of ^{18}F and ^{68}Ge are compared in Section 3.2.1. Differences in half-life and branching factors between are accounted for by applying the appropriate corrections when acquiring data. The mean positron range in water of ^{68}Ge (2.24mm) is longer than that of ^{18}F (0.64mm); whilst this may have implications for a small lesion study, this is highly unlikely to be of any disadvantage in the analysis of a relatively large uniform object.

The phantom's radioactivity concentration at the time of scanning was 3,631Bq/ml (±108Bq/ml) in the active volume (total activity 22.12MBq ± 0.66MBq). This total activity and activity concentration was similar to that recommended by GEMS for WCC using ^{18}F (approximately 20MBq into a 5,640ml volume, giving an approximate activity concentration of 3,546Bq/ml).

5.3.2 Image Acquisition and Reconstruction Protocols

The phantom was placed in the centre of the FOV. One single-bed, 4-minute PET acquisition was obtained. This phantom position ensured the central transaxial slice of the single bed acquisition (maximum sensitivity) coincided with the centre of the phantom. Correction factors for ^{68}Ge (half-life and branching factor) were applied to the acquired data.

In order to compare the quantitative accuracy of the WCC and clinical reconstructions, as per the first aim of this chapter, the phantom acquisition was reconstructed using the two sets of reconstruction parameters outlined in Table 5.1.

In order to characterise the effects of varying the reconstruction parameters, as per the second aim of this chapter, a wide range of parameters were examined. The transaxial FOV was fixed at 700mm, which is used for whole-body imaging. All possible settings for reconstruction method, matrix size and z-axis filter were assessed. The number of effective iterations and the Gaussian filter width, however, both have wide ranges of possible values, as described in Sections 3.1.2.3 and 3.1.2.5 respectively. A subset of the possible values for each of these parameters was chosen in order to effectively demonstrate the effect of each parameter, whilst keeping the size of the data set manageable. The Gaussian filter width was extended from 0mm to 10mm in steps of 1mm. The number of effective iterations ranged from 18 to 540. Both parameters were therefore extended beyond typical clinical settings to better evaluate their effects under extreme circumstances. Table 5.2 summarises the reconstruction parameters used.

Reconstruction Parameter	Settings Used	Number of Result Groups
Reconstruction Method	HD, PSF, TOF, PSF+TOF	4
Effective Iterations	18, 54, 90, 180, 540	5
Gaussian Filter (FWHM)	0 →10mm, in 1mm increments	11
Z-Axis Filter	None, Light, Standard, Heavy	4
Matrix Size	128*128, 192*192, 256*256	3

Table 5.2: Reconstruction parameters used for ^{68}Ge phantom

All possible combinations of the reconstruction parameters shown in Table 5.2 were used. 2,640 reconstructions were therefore performed in total.

5.3.3 Image Analysis

All ROI analysis was performed using Hermes Medical Systems' Hybrid Viewer software.

A large ROI (diameter 12cm) was positioned on the central PET slice and on two slices either side (5 ROIs in total). The mean and maximum voxel values were recorded for each ROI. The mean of the 5 mean ROI values was taken to be the mean activity concentration. The maximum voxel value across all 5 ROIs was taken to be the maximum activity concentration. This analysis was repeated for all reconstructions performed.

As all measurements were taken from the centre of the FOV, and not the FOV edges, quantitative values should be insensitive to filter settings according to the GEMS assertion being assessed by the second aim in this chapter.

5.4 Results

5.4.1 Comparison between WCC and Clinical Reconstructions

Figure 5.1 compares the central transaxial slices from the GEMS recommended WCC and clinical reconstructions, as defined in Table 5.1.

'GEMS WCC' Reconstruction 'GEMS Clinical' Reconstruction

Figure 5.1: Central transaxial slices of WCC and clinical reconstructions

Qualitative analysis of the 'GEMS WCC' and 'GEMS Clinical' reconstructions demonstrates little difference between the two images. Table 5.3 compares the mean and maximum voxel values for the 'GEMS WCC' and 'GEMS Clinical' reconstructions, and also compares them to the expected (reference) phantom activity concentration (3,631Bq/ml).

	GEMS WCC	GEMS Clinical
Mean Voxel Value	3,666Bq/ml	3,657Bq/ml
Difference from Reference	+35Bq/ml (+0.96%)	+26Bq/ml (+0.72%)
Maximum Voxel Value	5,132Bq/ml	5,074Bq/ml
Difference from Reference	+1,501Bq/ml (+41.3%)	+1,443Bq/ml (+39.74%)

Table 5.3: WCC and Clinical reconstruction voxel values

The difference between the mean voxel values for the 'GEMS WCC' and 'GEMS Clinical' images is small (9Bq/ml, 0.25% of expected activity concentration). The difference between the maximum voxel values is also small (58Bq/ml, 1.6% of expected activity concentration). The differences in quantitative voxel accuracy between the 'GEMS WCC' and 'GEMS Clinical' reconstructions for the 20cm uniform phantom are therefore highly unlikely to be of clinical significance.

The mean voxel values of the 'GEMS WCC' and 'GEMS Clinical' images are within 1% of the expected phantom activity concentration, and therefore within the phantom's ±3% tolerance. However, the maximum voxel values of both data sets are approximately 40% greater than the expected activity concentration. Maximum voxel values are 'single voxel' measurements and are therefore vulnerable to statistical noise; it is therefore expected that the maximum voxel values exceed the expected activity concentration.

The first aim of this chapter is therefore fulfilled: differences between the 'GEMS WCC' and 'GEMS Clinical' are minimal, and the current WCC protocol is therefore valid. The rest of this chapter will address the second aim of this chapter and characterise the effects of each reconstruction parameter upon voxel accuracy.

5.4.2 Effects of Reconstruction Parameters on Quantitative Accuracy

Before the effects of each reconstruction parameter are investigated in turn, the distribution of the full result set (all 2,640 reconstructions) for both mean and maximum activity concentrations are briefly examined. This is to inform an appropriate choice of statistical analyses when assessing the effects of each of the reconstruction parameters.

Figure 5.2 shows a histogram of all 2,640 mean activity concentration results, with a bin width of 1Bq/ml. The reference phantom activity and its ± 3% tolerance range are also shown. The histogram demonstrates all 2,640 mean activity concentration results are above the reference activity concentration, but within the stated ± 3% uncertainty of the calibrated phantom's activity concentration. The minimum result (3,647Bq/ml) is 0.44% above reference and the maximum result (3,700Bq/ml) is 1.9% above reference. The following factors may have contributed to this small positive offset, with respect to the reference value:

- Measurement error in the phantom's calibration at the time of manufacture
- Measurement error in the ^{18}F phantom used in the most recent WCC, used to generate the voxel values in the reconstructions performed in this chapter

A visual inspection of the histogram clearly demonstrates the mean activity concentration results from the full data set do not follow a normal distribution; this was confirmed by a Shapiro-Wilk test of normality (p-value <0.001). Non-parametric statistical tests are therefore used in this chapter to determine if there are any statistically significant differences in mean activity concentrations as a result of changes to the reconstruction algorithm. Furthermore, the full result set appears to demonstrate four distinct peaks. This is a result of the differences between the four reconstruction methods (HD, PSF, TOF and PSF+TOF), as will be demonstrated later in this chapter.

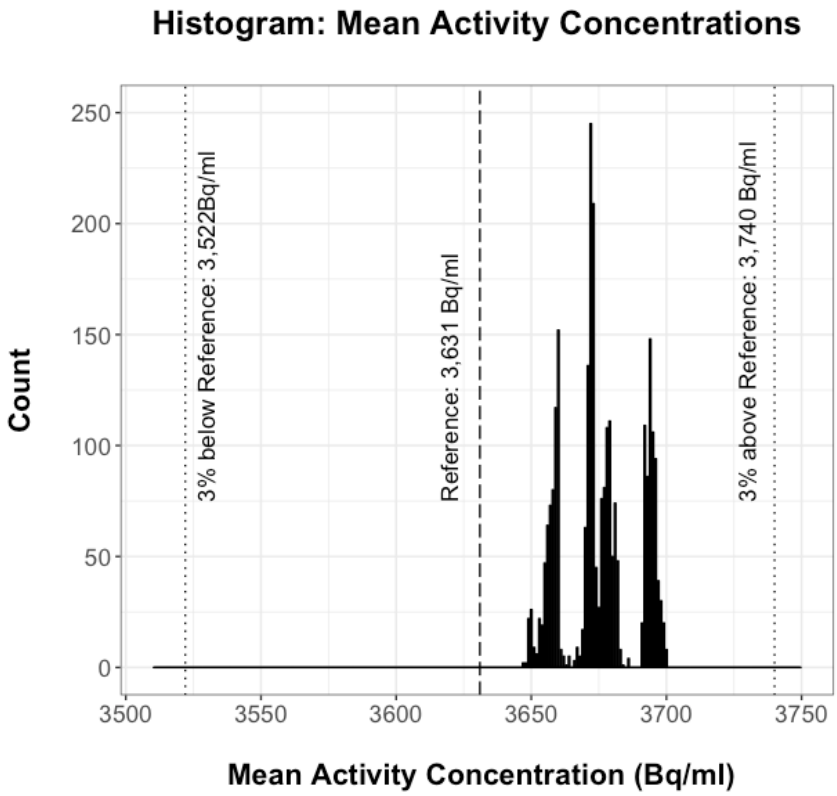


Figure 5.2: Histogram of all 2,640 mean activity concentration results Bin width is 1Bq/ml.

Figure 5.3 shows a histogram of all 2,640 maximum activity concentration results. The reference phantom activity and its ± 3% tolerance range are also shown. It should be noted that the y-axis range is much shorter than that of the mean activity concentrations, while the x-axis range is much greater (reference activity and tolerance limits shown at the extreme left of the graph).

Figure 5.3 demonstrates that all 2,640 maximum activity concentration results are above the upper bound of the phantom's calibrated activity concentration uncertainty. The minimum result (3,915Bq/ml) is 7.8% above reference and the maximum result, (27,300Bq/ml) is 652% above reference. As stated previously, maximum voxel values are vulnerable to statistical noise. The deviations from the reference phantom activity concentration depend on the statistical noise in the reconstructed image, which in turn depends on the reconstruction parameters used.

A visual inspection of maximum activity concentration histogram demonstrates that the results for the full data set do not demonstrate a normal distribution (Shapiro-Wilk p-value <0.001). Non-parametric statistical tests are therefore used for both mean and maximum activity concentration analysis in this chapter.

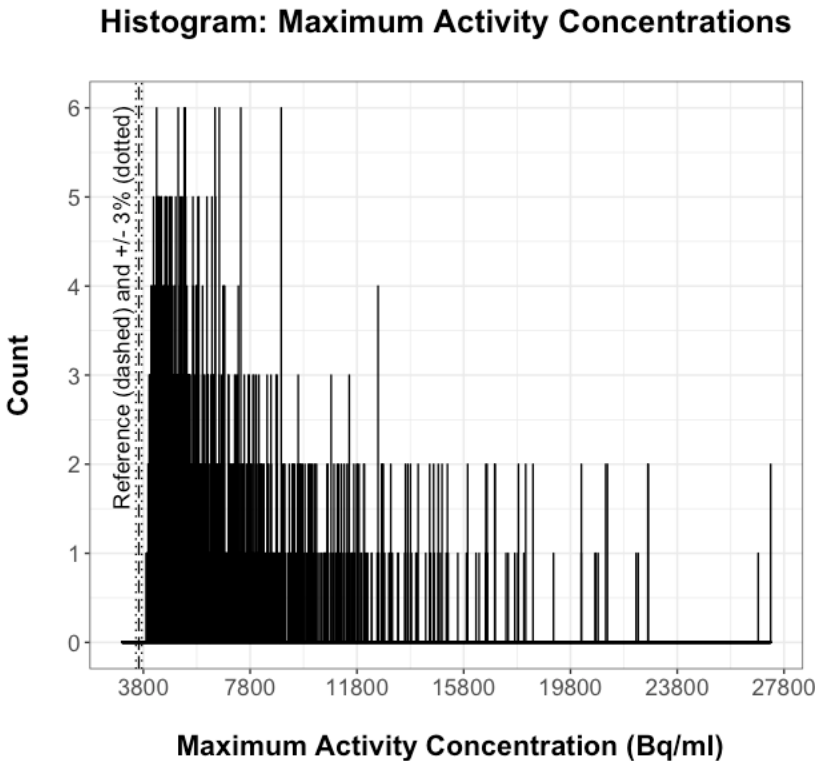


Figure 5.3: Histogram of all 2,640 maximum activity concentrations results Bin width is 1Bq/ml.

The remainder of the results section examines the effects of each of the reconstruction parameters in turn on both mean and maximum activity concentrations, in accordance with the second aim of this chapter. However, it should be noted that Figure 5.2 has already demonstrated the reconstruction parameters assessed in this chapter are unlikely to have any clinically significant impact upon mean activity concentrations in the uniform phantom, as the results for all 2,640 reconstructions are within 2% of the expected phantom activity concentration. Figure 5.3 demonstrates the reconstruction parameters do have a significant effect upon the maximum activity concentrations, as one would expect for 'single voxel' measurements that are vulnerable to image noise.

5.4.2.1 Effects of TOF and PSF

Figure 5.4 splits the mean activity concentration histogram data by reconstruction method. Comparing Figure 5.4 with Figure 5.2 confirms each of the four reconstruction methods accounts for a different peak in the histogram of the full result set.

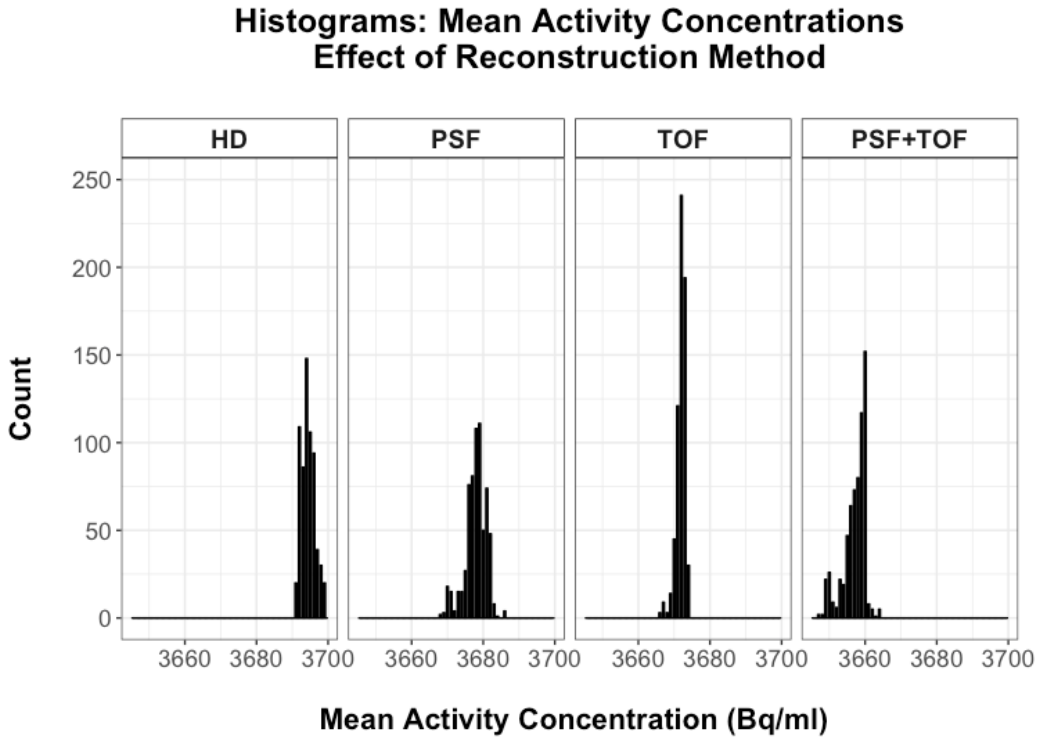


Figure 5.4: Mean activity concentrations histograms by reconstruction method Bin width is 1Bq/ml.

Figure 5.5 plots mean activity concentrations against reconstruction method in order to better visualise the differences between each method. The results are presented as box plots (shaded grey): the boxes show the first and third quartiles and the median value, the lower whiskers represent data within 1.5 of the inter-quartile range of the lower quartile, the upper whiskers represent data within 1.5 of the inter-quartile range of the upper quartile, and dots represent outliers. Violin plots (red lines) are overlaid onto the box plots and illustrate the distribution of results for each effective iteration setting.

Figure 5.5 demonstrates the reconstruction method has a clear effect on mean voxel values, in spite of the very small differences in activity concentrations. PSF-only appears to reduce voxel values when compared to HD. TOF-only reduces the voxel sizes below that of PSF-only. Applying both PSF and TOF together produces the lowest mean voxel values. However, all four median results are within 36Bq/ml (approximately 1%) of each other: 3,694Bq/ml (HD), 3,678Bq/ml (PSF), 3,672Bq/ml (TOF) and 3,658Bq/ml (PSF+TOF). The violin plot also demonstrates TOF-only produces the most consistent results, while the PSF reconstructions produce the least consistent results. This is confirmed by the differences between the maximum and minimum results for each method: TOF = 8Bq/ml; HD = 9Bq/ml; PSF+TOF = 17Bq/ml; PSF = 18Bq/ml.

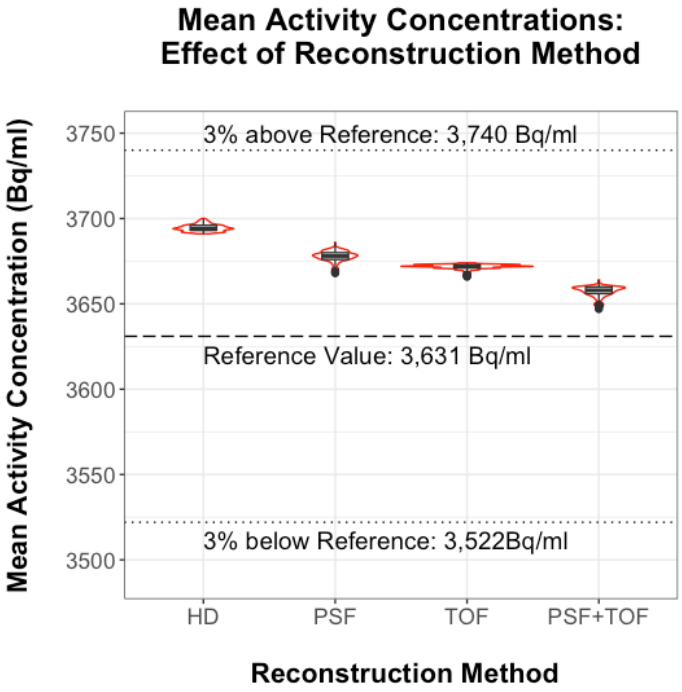


Figure 5.5: **Mean activity concentrations versus reconstruction method**

Results shown as box and violin plots. Whiskers in box-plots represent 1.5 inter-quartile ranges and dots represent outliers.

A Friedman's test confirmed there are statistically significant differences between the reconstruction methods (p-value <0.001). Pairwise Wilcoxon signed-ranks tests with Bonferroni multiple comparison corrections were therefore performed on each possible pair of reconstruction methods. The p-value in each case was <0.001, indicating each reconstruction method produces statistically significantly different results to the others. However, as stated previously, these differences are highly unlikely to be clinically significant given the small magnitude of differences in the activity concentrations.

Figure 5.6 plots maximum activity concentrations against reconstruction method. The results are presented as box and violin plots, in the same manner as Figure 5.5.

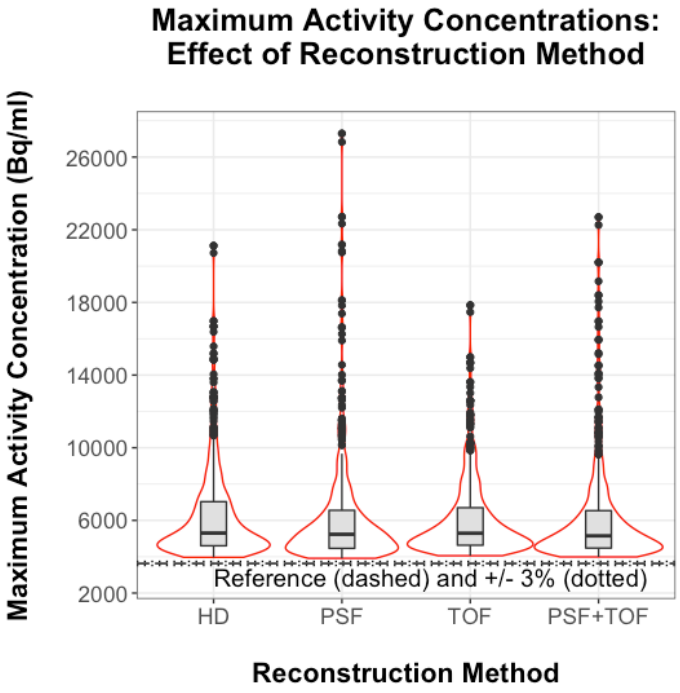


Figure 5.6: Maximum activity concentrations versus reconstruction method

Results shown as box and violin plots. Whiskers in box-plots represent 1.5 inter-quartile ranges and dots represent outliers.

Figure 5.6 demonstrates that reconstruction method has an effect on the outlying results. PSF produces the greatest outliers: maximum results were 27,300Bq/ml for PSF and 22,698Bq/ml for PSF+TOF. TOF-only produces the narrowest range of results: the maximum value for TOF (17,859Bq/ml) is less than that of the HD reconstruction (21,121Bq/ml). However, all four median results are within 174Bq/ml (approximately 3.3%) of each other: 5,304Bq/ml (HD), 5,326Bq/ml (PSF), 5,295Bq/ml (TOF) and 5,152Bq/ml (PSF+TOF).

The Friedman test returned a p-value of <0.001; Pairwise Wilcoxon signed-ranks (Bonferroni) tests were therefore performed. The p-value in each case was <0.05, indicating that each reconstruction method produces significantly different results to the others. Unlike the mean activity concentration results, these differences may be of clinical significance, particularly for the outliers.

5.4.2.2 Effects of Effective Iterations

Figure 5.7 plots mean activity concentrations against the number of effective iterations. Results are presented as box and violin plots.

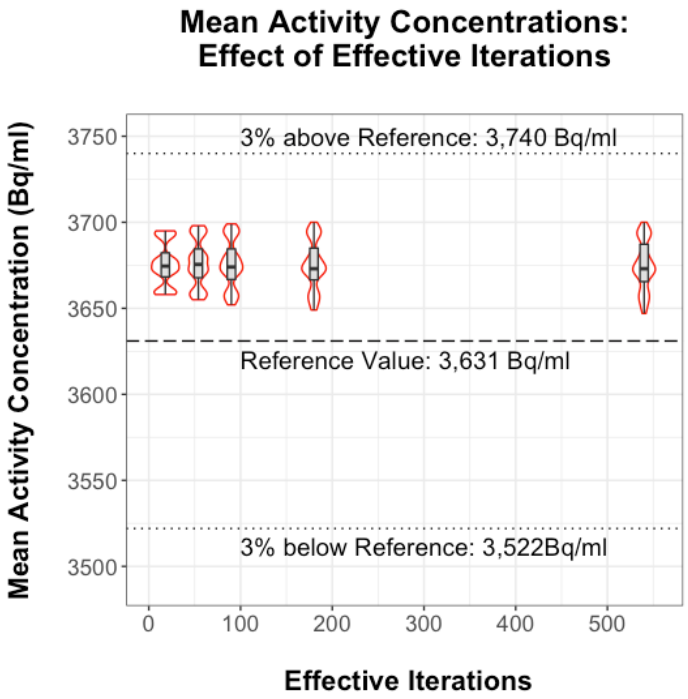


Figure 5.7: Mean activity concentrations versus effective iterations

Results shown as box and violin plots. Whiskers in box-plots represent maximum and minimum values (no outliers are present).

Figure 5.7 demonstrates that the range of results increases as more effective iterations are applied: the difference between the maximum and minimum result increases from 37Bq/ml (18 effective iterations) to 53Bq/ml (540 effective iterations). However, all 5 median results are within 2Bq/ml (0.05%) of each other (3,673Bq/ml for 180 and 540 iterations, 3,674Bq/ml for 18 and 90 iterations, and 3,675Bq/ml for 54 iterations).

The Friedman test returned a p-value of <0.001; Pairwise Wilcoxon signed-ranks (Bonferroni) tests were therefore performed. Only one result pair did not demonstrate a statistically significant difference (18 effective iterations versus 90 effective iterations returned a p-value of 0.789). However, as all results were within the narrow ranges discussed above, with very similar median results, these differences are highly unlikely to be clinically significant.

Figure 5.8 plots the maximum activity concentrations against the number of effective iterations. Results are presented as box and violin plots.

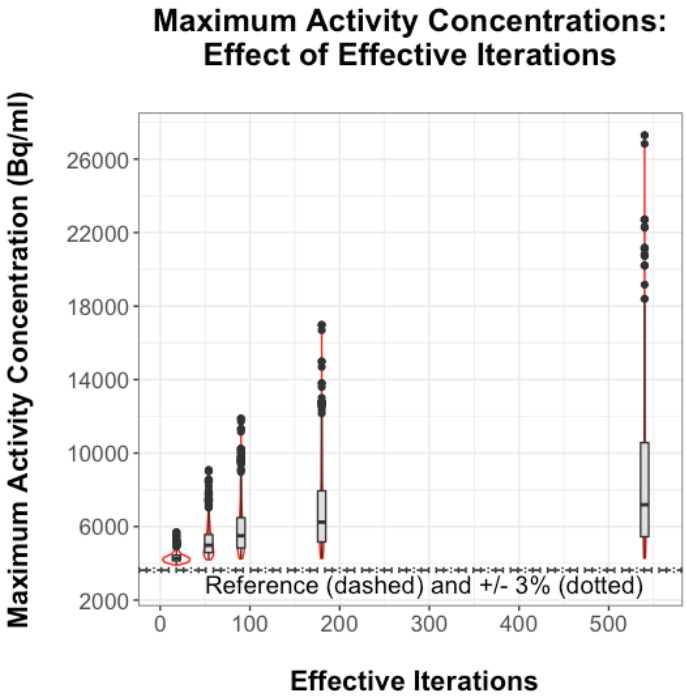


Figure 5.8: Maximum activity concentrations versus effective iterations

Results shown as box and violin plots. Whiskers in box-plots represent 1.5 inter-quartile ranges and dots represent outliers.

Figure 5.8 clearly demonstrates that as the number of effective iterations increases, the median result increases (from 4,269Bq/ml at 18 effective iterations to 7,189Bq/ml at 540 effective iterations). The ranges of results also increase with effective iterations. At 18 effective iterations, results range from 3,915Bq/ml to 5,695Bq/ml (difference of 1,780Bq/ml). At 540 effective iterations, results range from 4,259Bq/ml to 27,300Bq/ml (difference of 23,041Bq/ml).

The Friedman test returned a p-value of <0.001; Pairwise Wilcoxon signed-ranks (Bonferroni) tests were therefore performed. Each pair produced p-values <0.001, indicating each iteration setting produces statistically significantly different maximum results to the others. Unlike the mean voxel values, these differences are likely to be clinically significant, as the effect upon voxel values is large.

5.4.2.3 Effects of Matrix Size

Figure 5.9 plots mean activity concentrations against matrix size. Results are presented as box and violin plots.

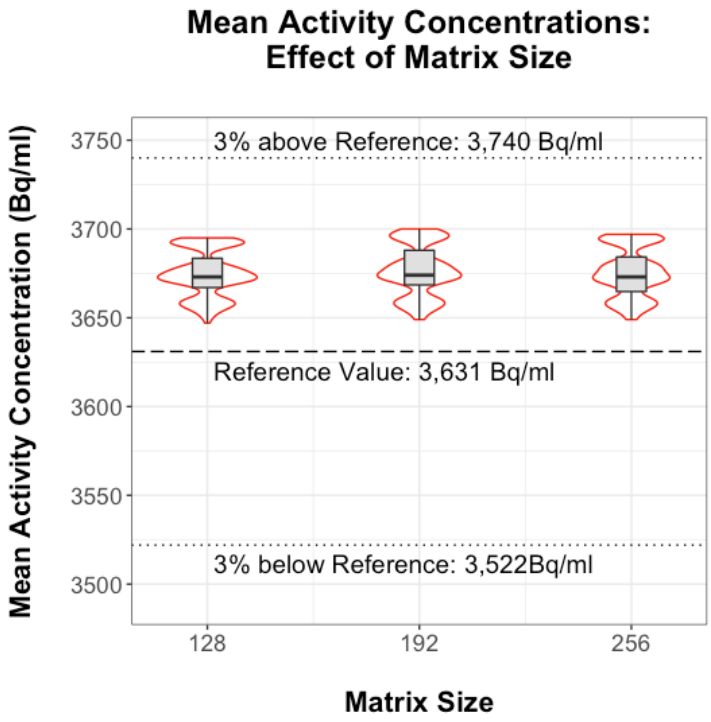


Figure 5.9: Mean activity concentrations versus matrix

Results shown as box and violin plots. Whiskers in box-plots represent maximum and minimum values (no outliers are present).

Figure 5.9 demonstrates that matrix size has little effect on mean voxel values. All three median values are within 1Bq/ml (0.03%) of each other: 3,673Bq/ml for the 128 and 256 matrices, and 3,674Bq/ml for the 192 matrix. The ranges of results for each matrix size are also very similar (48Bq/ml for both the 128 and 256 matrices and 51Bq/ml for the 192 matrix).

However, the Friedman test returned a p-value of < 0.001, indicating there are statistically significant differences in the mean voxel values caused by the choice of matrix size. This is not immediately obvious from Figure 5.9; however, upon closer inspection, the violin plots demonstrate different result distributions for each matrix setting. The 128 matrix distribution appears to be more focussed around the median result; however, as the matrix size increases, results appears to become more evenly distributed. Pairwise Wilcoxon signed-ranks (Bonferroni) tests produced p-values <0.001 for each pair, indicating each matrix size produces statistically significantly different results to the others. However, these differences are highly unlikely to be clinically significant, as the differences between the three sets of results are small.

Figure 5.10 plots maximum activity concentrations against matrix size. Results are presented as box and violin plots.

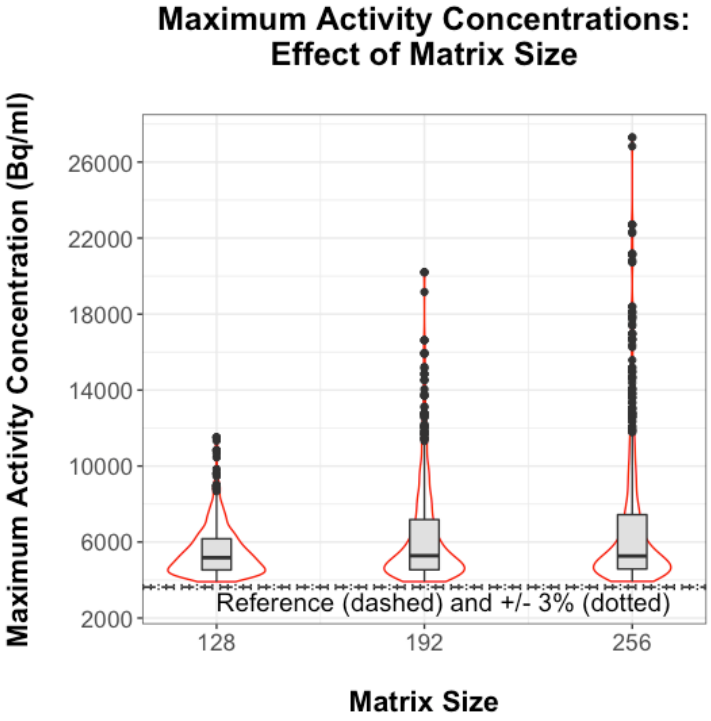


Figure 5.10: Maximum activity concentrations versus matrix size

Results shown as box and violin plots. Whiskers in box-plots represent 1.5 inter-quartile ranges and dots represent outliers.

In contrast to the mean activity concentration results, Figure 5.12 demonstrates that matrix size has a clear effect on the maximum activity concentrations. Median results for all three matrix sizes are within 105Bq/ml (2%) of each other: 5,176Bq/ml for the 128 matrix, 5,281Bq/ml for the 192 matrix and 5,263Bq/ml for the 256 matrix. However, the maximum outliers increase as the matrix size increases: 11,512Bq/ml (128 matrix), 20,209Bq/ml (192 matrix) and 27,300Bq/ml (256 matrix).

The Friedman test returned a p-value of <0.001; Pairwise Wilcoxon signed-ranks (Bonferroni) tests were therefore performed. Each pair produced p-values <0.001, indicating that each matrix size produces statistically significantly different results to the others. These differences may be of clinical significance as the effects upon voxel values are large.

5.4.2.4 Effects of Gaussian Filter Width

Figure 5.11 plots the mean activity concentrations against the post reconstruction filter FWHM.

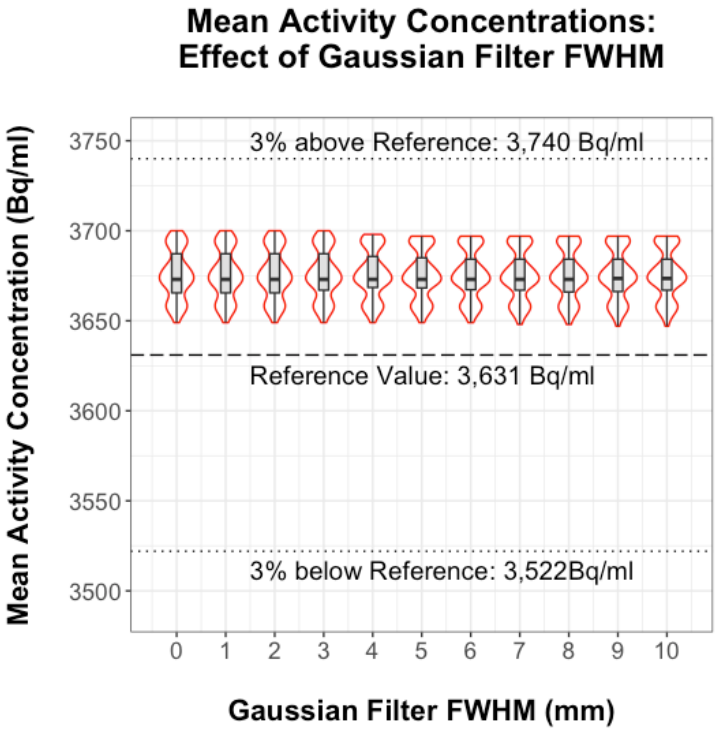


Figure 5.11: Mean activity concentrations versus Gaussian filter width

Results shown as box and violin plots. Whiskers in box-plots represent maximum and minimum values (no outliers are present).

The plots demonstrate that filter FWHM has very little effect on mean voxel values. The median result is almost identical for all eleven filter widths (3,673Bq/ml for filter widths 0mm through to 8mm, and 3674Bq/ml for filter widths 9mm and 10mm). The ranges of results are also almost identical (between 49Bq/ml and 51Bq/ml for all filter widths).

However, the Friedman test returned a p-value of <0.001, indicating there are statistically significant differences in mean voxel values caused by the filter FWHM. This is not immediately obvious from Figure 5.11; however, upon closer inspection, the violin plots demonstrate the distribution of results begin to change once the FWHM exceeds 3mm. As the filter width increases, results become increasingly distributed around the median result. Pairwise Wilcoxon signed-ranks (Bonferroni) tests were performed on each possible pair of filter widths (55 pairs possible from the eleven different filter widths). There are no statistically significant differences, at the 5% level of significance, between the following sets of results:

- 0mm, 1mm, 2mm and 3mm
- 6mm, 7mm and 8mm

All other result pairings demonstrated statistically significant differences. However, these differences are highly unlikely to be clinically significant, as the differences between the eleven result sets are small.

Figure 5.12 plots the maximum activity concentrations against the post reconstruction filter FWHM.

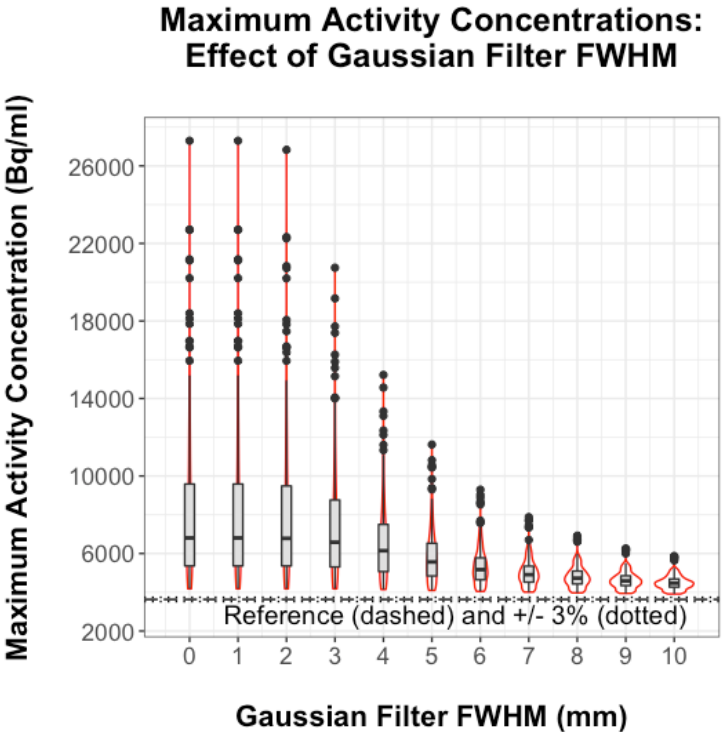


Figure 5.12: Maximum activity concentrations versus Gaussian filter width

Results shown as box and violin plots. Whiskers in box-plots represent 1.5 inter-quartile ranges and dots represent outliers.

Figure 5.12 demonstrates that filter FWHM has a clear effect on the maximum voxel values. The median value decreases with increasing filter width: from 6,803Bq/ml at FWHM = 0mm to 4,470Bq/ml at FWHM = 10mm. The result ranges also decrease with increasing filter width: from 23,120Bq/ml at 0mm to 1,959Bq/ml at FWHM = 10mm. Comparing Figure 5.11 with Figure 5.12 demonstrates that filter FWHM has a much greater effect on the maximum voxel values than the mean voxel values.

The Friedman test returned a p-value of <0.001; Pairwise Wilcoxon signed-ranks (Bonferroni) tests were therefore performed. The only results that did not demonstrate a statistically significant difference to each other were the results for 0mm and 1mm, whose distributions appear to be identical on Figure 5.12. All other pairs of results retuned p-values < 0.001. These differences are likely to be of clinical significance.

5.4.2.5 Effects of Z-Axis Filter Setting

Figure 5.13 plots the mean activity concentrations against the z-axis filter setting.

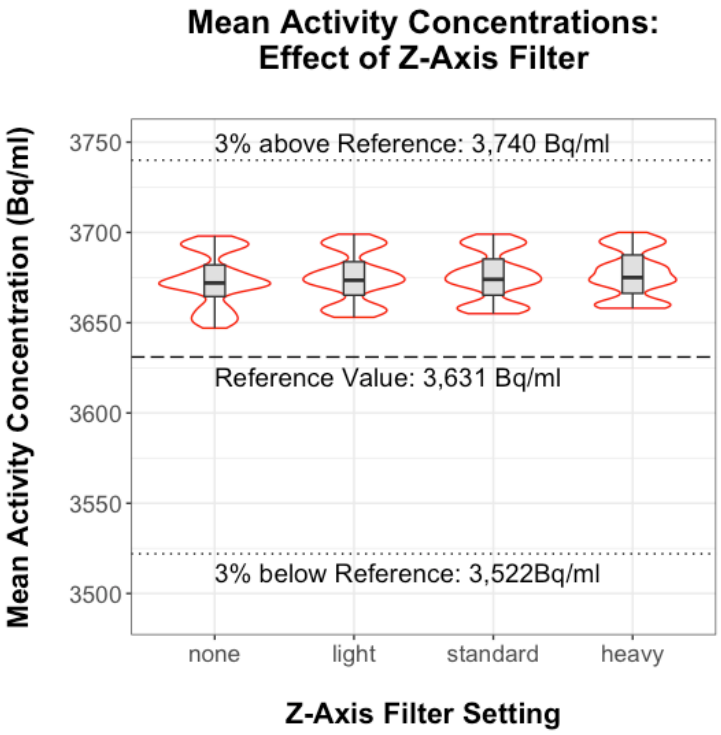


Figure 5.13: Mean activity concentrations versus z-axis filter

Results shown as box and violin plots. Whiskers in box-plots represent maximum and minimum values (no outliers are present).

Figure 5.13 demonstrates the z-axis filter has little effect on mean activity concentrations. Median results for all four filter weights are within 3Bq/ml (0.1%) of each other: 3,672Bq/ml (none), 3,673Bq/ml (light), 3,674Bq/ml (standard) and 3,675Bq/ml (heavy). However, the difference between maximum and minimum decreases as the filter weight increases: 51Bq/ml (none), 46Bq/ml (light), 44Bq/ml (standard) and 42Bq/ml (heavy).

The Friedman test returned a p-value of <0.001, indicating there are statistically significant differences in mean voxel values caused by the z-axis filter. This is not immediately obvious from Figure 5.13; however, upon closer inspection, the violin plots demonstrate that voxel values become increasingly distributed around the median result as the filter weight increases. Pairwise Wilcoxon signed-ranks (Bonferroni) tests were performed; the p-value in each case was <0.001, indicating each z-axis filter setting produces statistically significantly different results to the others. However, these differences are highly unlikely to be clinically significant, as the differences between the three sets of results are small.

Figure 5.14 plots the maximum activity concentrations against the z-axis filter setting.

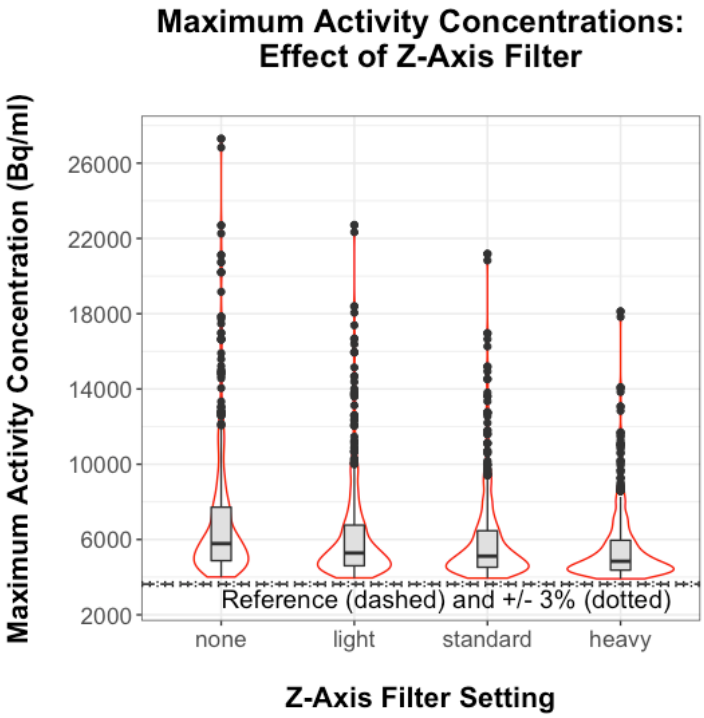


Figure 5.14: Maximum activity concentrations versus z-axis filter

Results shown as box and violin plots. Whiskers in box-plots represent 1.5 inter-quartile ranges and dots represent outliers.

Figure 5.14 demonstrates the z-axis filter does have an effect on maximum voxel values. Median results decrease as the weight of the z-axis filter increases: 5,784Bq/ml (none), 5,282Bq/ml (light), 5,114Bq/ml (standard) and 4,842Bq/ml (heavy). Maximum outliers also decrease as the z-axis filter weight increases: 27,300Bq/ml (none), 22,713Bq/ml (light), 21,182Bq/ml (standard) and 18,123Bq/ml (heavy).

The Friedman test returned a p-value of <0.001; Pairwise Wilcoxon signed-ranks (Bonferroni) tests were therefore performed. The p-value in each case was <0.001, indicating each z-axis filter setting produces statistically significantly different results to the others. These differences are likely to be of clinical significance.

5.5 Discussion

The first aim of this chapter was to assess the differences in the quantitative accuracy of uniform phantom images produced using the GEMS recommended WCC reconstruction parameters ('GEMS WCC') and the GEMS suggested reconstruction parameters for clinical imaging ('GEMS Clinical'). This was to test the GEMS assertion that the current Discovery 690 WCC protocol is suitable for use with the 'GEMS Clinical' reconstruction protocol. This assessment was made using a uniform cylindrical phantom with a similar size (20cm

diameter and 20cm length) and activity concentration (approximately 3,500Bq/ml) as the phantoms used to perform WCC.

Figure 5.1 and Table 5.3 demonstrated no clinically significant differences between the 'GEMS WCC' and 'GEMS Clinical' reconstructions, both qualitatively and quantitatively, for the uniform phantom. The mean voxels differed by 0.25% of the reference value, while the maximum voxels differed by 1.6%. As there were no clinically significant differences between the two reconstructions, it is therefore reasonable to conclude that performing WCC using either the 'GEMS WCC' or the 'GEMS Clinical' would produce clinically equivalent correction factors. The current GEMS protocol for the Discovery 690 WCC is therefore suitable for use with the 'GEMS Clinical' reconstruction parameters.

The second aim of this chapter was to determine if the quantitative accuracy of voxel values produced for a uniform phantom is insensitive to a wider range of reconstruction parameters, as stated by GEMS. In order to do this, the effects of each reconstruction parameter on both mean and maximum voxel values were characterised in turn. The effects of each of the reconstruction parameters upon both mean and maximum voxel values were largely as one would expect, given their known effects upon reconstructed images (described in Chapters 1 and 2), and are summarised in the following paragraphs.

Although there were statistically significant differences in mean voxel values as a result of altering each of the five reconstruction variables, the magnitudes of the differences were small in each case: mean voxel results ranged from 3,647Bq/ml (0.4% above phantom reference) to 3,700Bq/ml (1.9% above phantom reference). All mean activity concentration measurements were therefore within the phantom's ±3% tolerance. These differences are therefore highly unlikely to be of any clinical significance. However, the analysis did highlight the following effects:

- Increasing the number of effective iterations reduced the consistency of mean voxel values, as a result of increased image noise (Figure 5.7)
- Increasing the matrix size, and therefore decreasing the voxel size, marginally reduced the consistency of mean voxel values, as a result of increased image noise (Figure 5.9)
- Increasing the Gaussian filter FWHM marginally improved the consistency of mean voxel values as a result of reduced image noise (Figure 5.11)
- Increasing the weight of the z-axis filter marginally improved the consistency of mean voxel values, as a result of reduced image noise (Figure 5.13)
- Applying PSF reduced mean voxel values when compared with the HD reconstruction, which were closer to the reference value, and produced the least consistent results (Figure 5.5)

- Applying TOF reduced the mean voxel values compared to both HD and PSF, which were closer to the reference value, and produced the most consistent results of all four reconstruction methods (Figure 5.5)
- Applying both PSF and TOF together produced mean voxels closer to the reference value than the other three reconstruction methods (Figure 5.5)

Most of the above effects are as one would expect. Increasing the number of effective iterations, reducing voxel sizes and reducing the degree of filtering are all known to increase noise within the reconstructed images, as discussed in Chapters 1 and 2. Noise has a greater effect on maximum voxels than mean voxels; however, these results demonstrate corresponding measurable, but clinically insignificant, effects on mean voxels. The effects of PSF and TOF are more challenging to interpret. One would expect both of these corrections to reduce image noise when compared to the HD reconstruction. This may explain why both PSF and TOF produced mean voxel values which were closer to the reference value. However, as all mean voxel results were within the stated ± 3% uncertainty of the calibrated phantom's activity concentration, it is impossible to say which reconstruction produced the most accurate results. It may also be counter-intuitive that PSF produced less consistent results than HD, which may indicate greater noise levels than HD; this is addressed later in this chapter.

Overall, the mean voxel results are consistent with the GEMS assertion regarding the insensitivity of quantitative values for large, uniform objects to the reconstruction parameters.

Analysis of the maximum voxel values, however, demonstrated the effects of altering each of the reconstruction parameters are not only statistically significant, but are likely to be clinically significant, as the magnitude of the effects were large in relation to the reference voxel value. The following effects were observed:

- Increasing the number of effective iterations increased both the range, and the median, of the maximum voxel results, as a result of increased image noise (Figure 5.8)
- Increasing the matrix size, and therefore decreasing the voxel size, increased the range of maximum voxel results, as a result of increased image noise; however, the median results remained similar for all three matrix sizes. The main differences between the three matrix sizes were with the outlying results. (Figure 5.10)
- Increasing the Gaussian filter FWHM decreased both the range, and the median, of the maximum voxel results (Figure 5.12)
- Increasing the weight of the z-axis filter decreased the range, and the median, of the maximum voxel results (Figure 5.14)

- Applying PSF increased the range of maximum voxel results, while TOF reduced the range of maximum voxel results. However, the median results remained similar for all four reconstruction methods. The main differences between the four reconstruction methods are with the outlying results (Figure 5.6)

As with the mean results, the application of PSF appears to have increased image noise for both mean and maximum voxel values; however, one would expect both PSF and TOF to reduce image noise, as discussed in Chapters 1 and 2. As the performance of both PSF and TOF are dependent upon effective iterations, a second plot of maximum voxel values against reconstruction method is now presented in Figure 5.15: effective iterations have been restricted to the typical clinical range (the results for 180 and 540 effective iterations have been removed from the dataset).

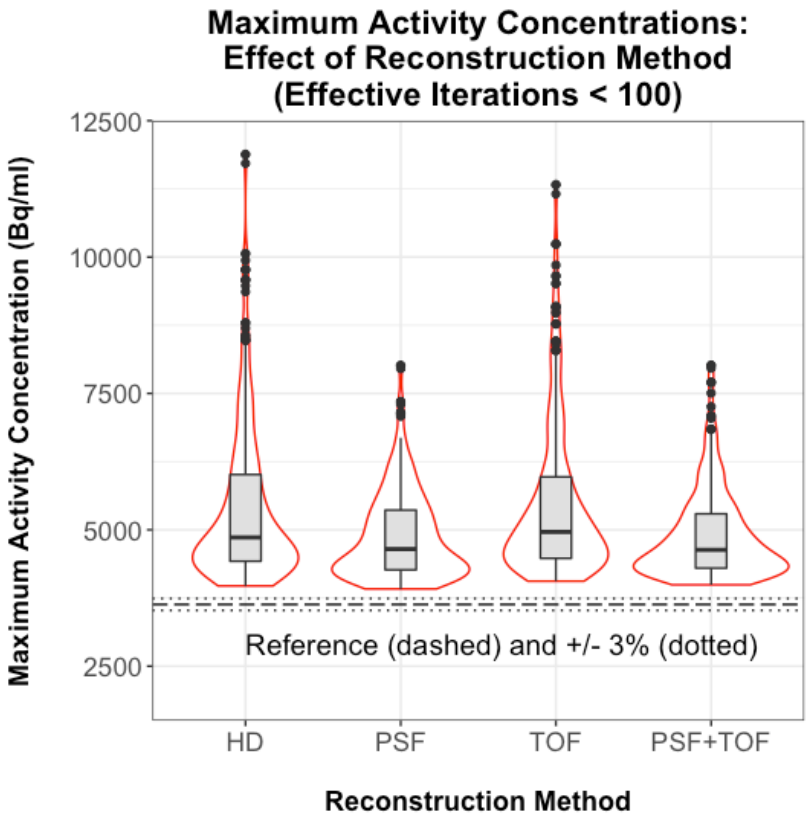


Figure 5.15: Maximum activity concentrations versus reconstruction method (effective iterations restricted to < 100)

Results shown as box and violin plots. Whiskers in box-plots represent 1.5 inter-quartile ranges and dots represent outliers.

Figure 5.15 demonstrates that when effective iterations are limited to <100, TOF continues to produce more consistent results than HD; however, PSF now produces the most consistent maximum voxel values. Figures 5.6 and 5.15 demonstrate that PSF is particularly vulnerable to noise when large numbers of effective iterations are applied but reduces noise

at clinically relevant levels of effective iterations. The effects of reconstruction parameters on image noise are investigated in more detail in the next chapter.

Maximum voxel results are therefore not consistent with the GEMS statement regarding the insensitivity of quantitative values for large, uniform objects to the reconstruction parameters.

With regards to WCC, one can infer from the personal communications with GEMS, combined with the results in this chapter, that only the mean voxel values from the WCC reconstruction are applicable to the correction factors used for quantitative PET data.

With regards to the interpretation of clinical data, the effects of reconstruction parameters upon maximum voxel values are likely to be clinically significant. SUV_{max}, as described in Section 1.5.2, is calculated using the maximum voxel within a ROI, and remains widely used in clinical PET. However, as demonstrated in this chapter, any measurement based upon maximum voxels are vulnerable to noise introduced by the reconstruction process. The effects of reconstruction parameters upon clinical data interpretation are investigated in more detail later in this thesis.

Finally, this chapter assessed the effects of reconstruction parameters using a single phantom: a 20cm diameter cylindrical uniform phantom. One would expect similar results for mean voxel values with any uniform phantom large enough to prevent partial volume effects; however, larger phantoms would be inherently noisier due to increased scatter and attenuation. Maximum voxel values would therefore likely increase as the size of the phantom increases. The ability of TOF to restrict noise, and therefore improve the accuracy of voxel values, would be better demonstrated by larger phantoms, as discussed in Section 1.7.1. Furthermore, it is of interest that mean voxel values within the 20cm uniform phantom reached convergence after only 18 effective iterations. Smaller objects within a non-uniform phantom would require more iterations to converge to accurate voxel values. Non-uniform phantoms, by definition, would not be used for WCC purposes, as a mean voxel value corresponding to a known activity concentration is required to generate the WCC correction factors. The convergence of smaller objects is assessed in Chapter 9 of this thesis.

A limitation of this study was that a single acquisition of the ^{68}Ge phantom was used for analysis. The methodology could be repeated multiple times using a ^{18}F phantom to demonstrate the results in this chapter are reproducible. However, one would not expect the overall findings of this chapter to be altered by the repeated measurements, as the observed effects of each reconstruction parameter were as expected.

5.6 Conclusions

This chapter concludes that the WCC reconstruction methodology employed by GEMS for the Discovery 690 is compatible with their suggested clinical reconstruction parameters.

Furthermore, this chapter concludes that the GEMS assertion regarding the insensitivity of quantitative values of large, uniform objects to reconstruction parameters is valid for mean voxel values, for the range of reconstruction parameters assessed in this chapter when a 20cm cylindrical phantom is used. One would expect this assertion to hold for any uniform object large enough to prevent partial volume effects.

Further work in this area could involve a more direct assessment of the effects of reconstruction parameters upon WCC. WCC calibrations could be performed using a range of different reconstructions. Correction factors generated by each reconstruction could be recorded and compared. Each set of WCC correction factors could be applied to phantom acquisitions, followed by analysis of the effects on quantitative voxel values. However, experimentation involving WCC factors on clinical PET-CT systems should be conducted with care to prevent any inappropriate correction factors being applied to clinical scans.

Chapter 6 : Assessment of Image Noise in Reconstructed Images of a Uniform Phantom

The previous chapter assessed the effects of varying reconstruction parameters upon voxel accuracy within a uniform phantom. The chapter concluded that mean voxel values were insensitive to changes in reconstruction parameters, as per the GEMS assertion. However, maximum voxel values, which are 'single voxel' measurements and therefore vulnerable to statistical noise, were shown to be significantly affected by reconstruction parameters. Whilst maximum voxel values reflect the effects of noise upon voxel accuracy, they cannot be used to directly quantify image noise itself.

Noise Equivalent Count Rate (NECR) tests are performed to NEMA standards during acceptance testing of new PET-CT systems (and are typically performed on an annual basis thereafter). These tests are well known in the field and provide a global noise assessment that allows PET system performance to be compared against manufacturer specifications and other PET systems. Some studies have performed NECR-style measurements from patient data when assessing the effects of reducing injected activities or increasing patient body mass index (BMI) [185], [186] (for example, using the DICOM image header to extract true, random and scatter data [185]). However, NECR is a raw data metric that does not account for variations within the FOV or the effects of reconstruction algorithms [187].

As discussed in Chapter 2, reconstructed image noise is typically assessed using voxel-to-voxel noise metrics (such as image roughness and COV) and region-to-region metrics (such as background variation, used by NEMA image quality assessment). Voxel-to-voxel noise metrics are generally taken to represent the noise perceived when viewing an individual image, while region-to-region metrics assess non-uniformities throughout the image [90].

Chapter 2 also demonstrated that several studies have assessed the effects of various reconstruction parameters upon both voxel-to-voxel and region-to-region noise. However, to the authors knowledge, no study has fully assessed the combined effects of effective iterations, filtering and matrix sizes when used with PSF and/or TOF. Furthermore, there is no universally agreed approach to clinically relevant noise assessment in PET imaging.

The primary aim of this thesis is to develop a generic methodology to assess and optimise PET image reconstruction whilst considering clinical context. This chapter uses the large reconstruction set produced in Chapter 5 to assess the effects of the various reconstruction parameters upon image noise. The use of the widely available ^{68}Ge cylindrical uniform phantom will therefore be investigated for clinically relevant noise assessment.

In order to optimise PET reconstruction on the GEMS Discovery 690 PET-CT system with respect to the detection of liver lesions, and therefore fulfil the secondary aim of this thesis, it is important to fully characterise the effects various reconstruction parameters have upon image noise. When combined with assessments of spatial resolution and lesion detection, which are explored in subsequent chapters, this will enable the selection of a combination of reconstruction parameters that produce the optimal trade-off between image noise and other aspects of image quality, in order to maximise lesion detection within the liver.

6.1 Introduction

Different methods for measuring PET image noise have been discussed in the literature. The NEMA Standards Publication for PET performance measurements [70] describes the use of Background Variability. This metric measures region-to-region variability within the image. Twelve non-overlapping ROIs of diameter 3.7cm are placed within the uniform background activity on the central slice of the NEMA torso phantom. These spheres are copied to 2 slices either side of the central slice, giving 60 ROIs in total. The mean voxel value for each of the 60 ROIs are recorded. Background Variability is then defined as the COV of these 60 ROI means (see Section 3.4 for full equation). This method has been widely used in the literature for assessing noise in PET phantom studies [19], [43], [90], [92], [188].

Image Roughness measures voxel-to-voxel variability within a PET image and is similar to the basic COV method. Its use in the literature is often based upon the same 60 ROIs described by NEMA Background Variation [43], [90], [92], [146], although larger/fewer ROIs have also been used [134], [189]. A COV is calculated for each of the 60 ROIs, and Image Roughness is taken to be the average of the 60 COVs (see Section 3.4 for full equation).

The use of Image Roughness is also common in ^{18}F-FDG patient studies, using a small number of ROIs placed over the relatively uniform distribution within the liver [189]–[191].

Both Background Variability and Image Roughness can be calculated for a single realisation/acquisition or for multiple realisations/acquisitions [92]. The use of multiple realisations has often been considered to be more closely related to the true noise in an image because single realisation measurements can be affected by correlations between adjacent voxels [189]. However, multiple realisations are not practical for patient studies due to the relatively low clinical acquisition times: rebinning clinically acquired patient studies into multiple realisations would result in unrealistically low-count, high noise images.

The use of single realisation measurements has therefore been justified by previous studies by comparing measurements of long-lived phantoms using both single realisation and multiple realisations. Studies by Tong et al [90], [92] concluded that NEMA Background Variation using a single realisation is a reasonable surrogate for noise measurements using multiple realisations. Mettivier et al [43] compared the use of multiple slices within a single realisation with the use of multiple realisations and concluded that Image Roughness analysis using multiple slices within a uniform phantom was comparable to that of using multiple realisations.

The effects of effective OSEM iterations, filtering, voxel size, PSF and TOF upon PET image noise have been widely assessed in the literature, as described in Chapter 2. However, to the authors' knowledge, no single study has examined the effects of all of these parameters in combination. In particular, the combined effect of matrix size and filters on PET image noise has not yet been evaluated fully. Furthermore, there is no universally agreed approach to clinically relevant noise assessment in PET imaging.

6.2 Aims

This chapter aims to characterise the combined effects of the following reconstruction parameters on both Image Roughness and Background Variation within images of a 20cm diameter uniform ^{68}Ge phantom:

a. Effective iterations
b. Gaussian Filter Width
c. Z-Axis Filter Setting
d. Matrix size (and hence voxel size)
e. Reconstruction method (TOF and PSF)

In particular, this chapter aims to examine the combined effect of matrix size and filters on image noise, with a view to assessing how the use of a 256 matrix (instead of the GEMS suggested 192 matrix) to improve image resolution may impact the noise in the images.

This chapter will also discuss the effectiveness of the 20cm diameter uniform ^{68}Ge phantom for clinically relevant noise assessment.

Results will also be discussed with respect to the European guidelines for reconstruction parameters, as discussed in Section 2.1.

6.3 Materials and Methods

All acquisitions and reconstructions were performed on the GEMS Discovery 690 PET-CT system. A detailed description of this system is included in Chapter 3.

6.3.1 Phantom

This experiment uses the same uniform ^{68}Ge 20cm diameter cylindrical phantom as Chapter 5. The phantom's radioactivity concentration was 3,631Bq/ml at the time of scanning. This provides a conservative estimate of 400MBq ^{18}F-FDG patient liver count density, as previously described in Chapter 4.

6.3.2 Image Acquisition and Reconstruction Protocols

This chapter uses the same single-bed acquisition and 2,640 reconstructions as Chapter 5. The reconstruction parameter settings are shown in Table 6.1.

Reconstruction Parameter	Settings Used	Number of Result Groups
Reconstruction Method	HD, PSF, TOF, PSF+TOF	5
Effective Iterations	18, 54, 90, 180, 540	4
Gaussian Filter (FWHM)	0 →10mm, in 1mm increments	11
Z-Axis Filter	None, Light, Standard, Heavy	4
Matrix Size	128*128, 192*192, 256*256	3

Table 6.1: Reconstruction parameters used for ^{68}Ge phantom

Although 540 effective iterations are unlikely to be used clinically and were shown to significantly affect maximum voxel values in the previous chapter, several studies have shown that PSF requires more than 180 effective iterations to reach full convergence for

small objects when the GEMS Discovery 690 PET-CT system is used (discussed in Chapter 2). It is therefore prudent to include 540 effective iterations in this chapter, as high numbers of effective iterations may be relevant to later chapters investigating spatial resolution and contrast recovery.

6.3.3 Image Analysis

All ROI analysis was performed using Hermes Medical Systems' Hybrid Viewer software. Twelve non-overlapping ROIs of diameter 3.7cm were positioned on the central PET slice and on two slices either side (60 ROIs in total). ROI placement for the phantom's central slice is illustrated in Figure 6.1. The mean voxel value and standard deviation were recorded for each of the 60 ROIs. The following metrics were then calculated for each reconstruction:

- Image Roughness (IR): COV was calculated for each of the 60 individual ROIs and then averaged over the 60 ROIs to produce the IR measurement (Equation 3.3)
- Background Variability (BV): COV of the 60 ROI mean values was calculated to produce the BV measurement (Equation 3.4).

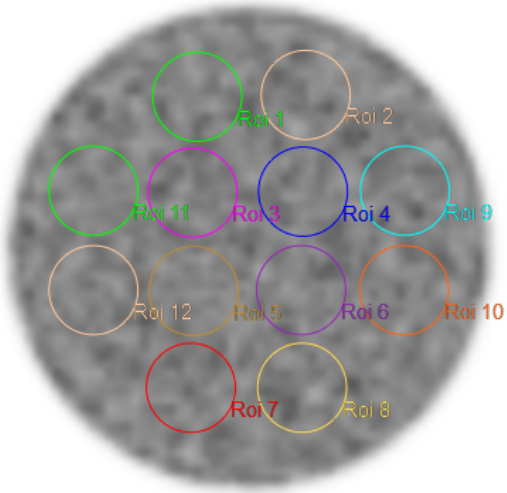


Figure 6.1: ROI Placement on central slice of uniform phantom

The EANM guidelines refer to the EARL procedure for assessing PET-CT system specific patient FDG activities with respect to image quality [121]. Both the EARL and QIBA publications specify that image noise, as measured by the COV in a volume of uniform activity, should not exceed 15%. As Image Roughness is similar to COV (Image Roughness is the average of multiple ROI COVs), Image Roughness results are compared to this 15% noise limit. As the phantom count density is lower than that observed in ^{18}F-FDG patient livers, this chapter provides a conservative assessment of clinical image quality.

6.4 Results

The first part of the results section assesses the impact of each reconstruction parameter in turn upon both Image Roughness and Background Variation. The second part of the results section explores an unexpected relationship between image noise, matrix size and Gaussian filter width that has not yet been discussed in the literature.

6.4.1 Impact of Reconstruction Parameters on Image Noise

6.4.1.1 Effects of Effective Iterations

Figure 6.2 illustrates the effects of effective iterations on both Image Roughness and Background Variation for the full result set (all 2,640 results are included in Figure 6.2). The results are presented as box and violin plots, in the same manner as the results in the previous chapter. Each box plot is annotated with the median value.

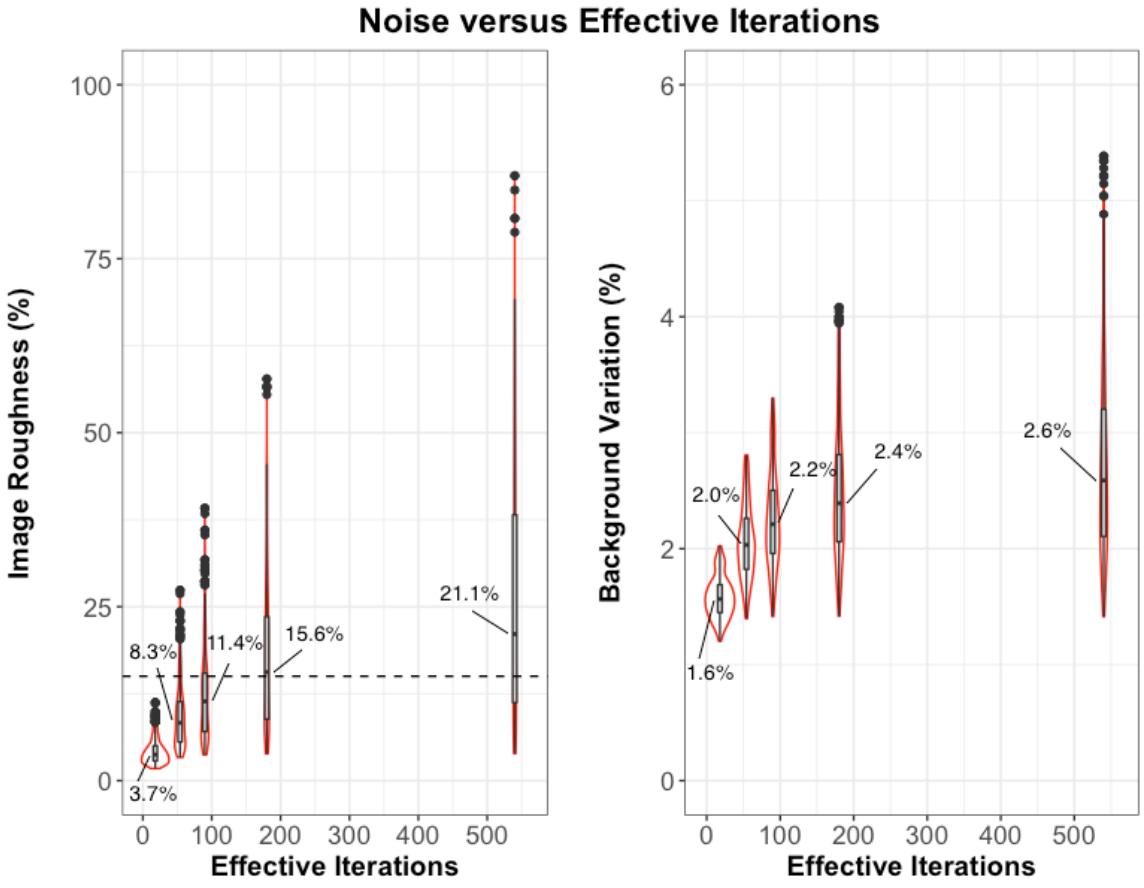


Figure 6.2: Image Roughness and Background Variation versus effective iterations

Results shown as box and violin plots. Whiskers in box-plots represent 1.5 inter-quartile ranges and dots represent outliers. Dashed line on Image Roughness plot represents 15% COV limit suggested by guidelines. Annotations show median result for each iteration setting.

Figure 6.2 clearly demonstrates that as the number of effective iterations increases, the median results and outliers for Image Roughness increase: 540 effective iterations produce a median result almost six times greater than that of 18 effective iterations, and a maximum result over 7 times greater. This is as one would expect, as image noise is known to increase as the number of effective iterations are increased [43], (discussed in Chapters 1 and 2). The Image Roughness results also mirror those of the maximum voxels in the previous chapter (Figure 5.8). All results at 18 effective iterations are below the 15% EARL limit, while the median results for 54 and 90 effective iterations are below 15%. The median result for 180 effective iterations slightly exceeds the limit (15.6%), while the majority of results for 540 effective iterations exceed the limit.

Figure 6.2 also demonstrates that Background Variation results increase as the number of effective iterations increase; however, these increases are to a lesser degree than those observed for Image Roughness. 540 effective iterations produce a median result less than double that of 18 effective iterations, and a maximum result less than 3 times greater.

Statistical tests were performed on both the Image Roughness and Background Variation results. In both cases, the Friedman test returned a p-value of < 0.001, confirming statistically significant differences in both Image Roughness and Background Variation are caused by the number of effective iterations. Pairwise Wilcoxon signed-ranks (Bonferroni) tests were then performed on each possible pair of effective iterations settings. The p-value in each case was < 0.001, indicating that each effective iteration setting produces statistically significantly different Image Roughness and Background Variation results to the others at the 1% level of significance.

The differences in both noise metrics caused by increasing effective iterations could be of clinical significance, as the effects upon each metric's median values are proportionally large. The increase in noise may affect observer confidence in reporting clinical images, or make it increasingly difficult to visualise small, low intensity lesions within a noisy background activity. Furthermore, using more than 200 effective iterations is likely to produce images with Image Roughness exceeding the 15% EANM guideline, depending on the particular combination of parameters used.

Figure 6.2 demonstrates that effective iterations have a significant effect upon image noise. In order to prevent the effects of effective iterations from masking the subtler effects of the other reconstruction parameters, the remainder of this section splits each result set by effective iterations.

6.4.1.2 Effects of Gaussian Filter Width

Figure 6.3 illustrates the effects of the Gaussian filter FWHM on both Image Roughness and Background Variation for the full result set. Results are presented as box and violin plots. Each row represents a different number of effective iterations. Different y-axis limits are used for each row to better illustrate the effects of the Gaussian filter at each iteration setting.

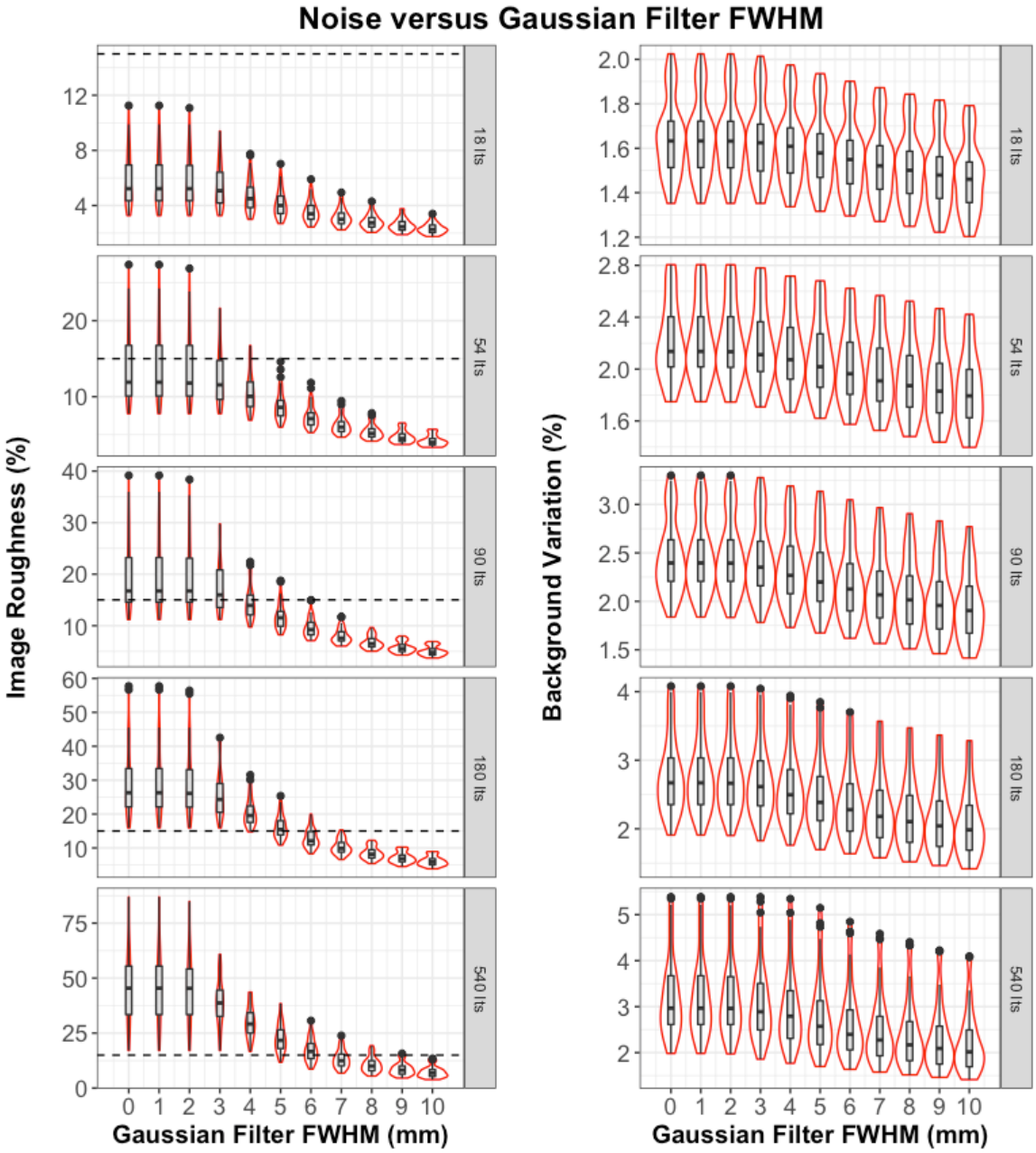


Figure 6.3: Image Roughness and Background Variation versus Gaussian filter

Results shown as box and violin plots. Whiskers in box-plots represent 1.5 inter-quartile ranges and dots represent outliers. Each row represents a different number of effective iterations, with different y-axis limits. Dashed lines on Image Roughness plot represents 15% COV limit suggested by guidelines.

Figure 6.3 clearly demonstrates that as the Gaussian filter width increases, both the median and maximum results for Image Roughness decrease. This is as one would expect, as filtering is known to reduce image noise [44] (discussed in Chapter 1); however, it is of interest that the results for the 0mm and 1mm filters are identical. Figure 6.3 also demonstrates that the proportional decrease in Image Roughness as a result of filtering increases as the number of effective iterations increases. For example, when 18 effective iterations are applied (top row), the 10mm filter median Image Roughness is 43% of that of the 0mm filter. When 540 effective iterations are applied (bottom row), the 10mm filter median Image Roughness is only 15% of that of the 0mm filter. Figure 6.3 also demonstrates which filter settings are required to achieve the 15% EARL noise limit for each effective iteration setting. All results for 18 effective iterations are below the threshold. When 54 effective iterations are applied, all filter settings produce median results below the threshold; however, a filter width of at least 5mm is required to prevent any outlying results breaching the threshold. A 4mm filter is required for median results to achieve the 15% threshold when 90 effective iterations are applied, rising to 6mm for 180 effective iterations and 7mm for 540 effective iterations.

Figure 6.3 also demonstrates that Background Variation results decrease as the Gaussian filter width increases; however, as with the effective iterations results, the decreases in Background Variation as a result of the Gaussian filter are to a lesser degree than those observed for Image Roughness. For example, when 18 effective iterations are applied, the 10mm filter median Background Variation is 89% of that of the 0mm filter. When 540 effective iterations are applied, the 10mm filter median Background Variation is 68% of that of the 0mm filter.

The Friedman test returned a p-value of < 0.001 for both Image Roughness and Background Variation; pairwise Wilcoxon signed-ranks (Bonferroni) tests were therefore performed on each possible pair of filter widths at each iteration setting. As expected, no significant differences were found between the 0mm and 1mm filters for either Image Roughness or Background Variation. However, all other filter widths were found to be statistically significantly different to each other ($p < 0.01$ in each case). These differences may be of clinical significance. For example, an image produced using a 10mm Gaussian filter will have a fraction of the noise found in an image produced without any Gaussian filtering: this is highly likely to affect image interpretation. However, while images produced using 1mm and 2mm filters demonstrated statistically significant differences in median values, their differences are highly unlikely to be clinically significant due to their small magnitude.

6.4.1.3 Effects of Z-Axis Filter

Figure 6.4 illustrates the effects of the Z-axis filter setting on both Image Roughness and Background Variation for the full result set. Results are presented as box and violin plots. Each row represents a different number of effective iterations. Different y-axis limits are used for each row to better illustrate the effects of the z-axis filter at each iteration setting. Each box plot is annotated with the median value.

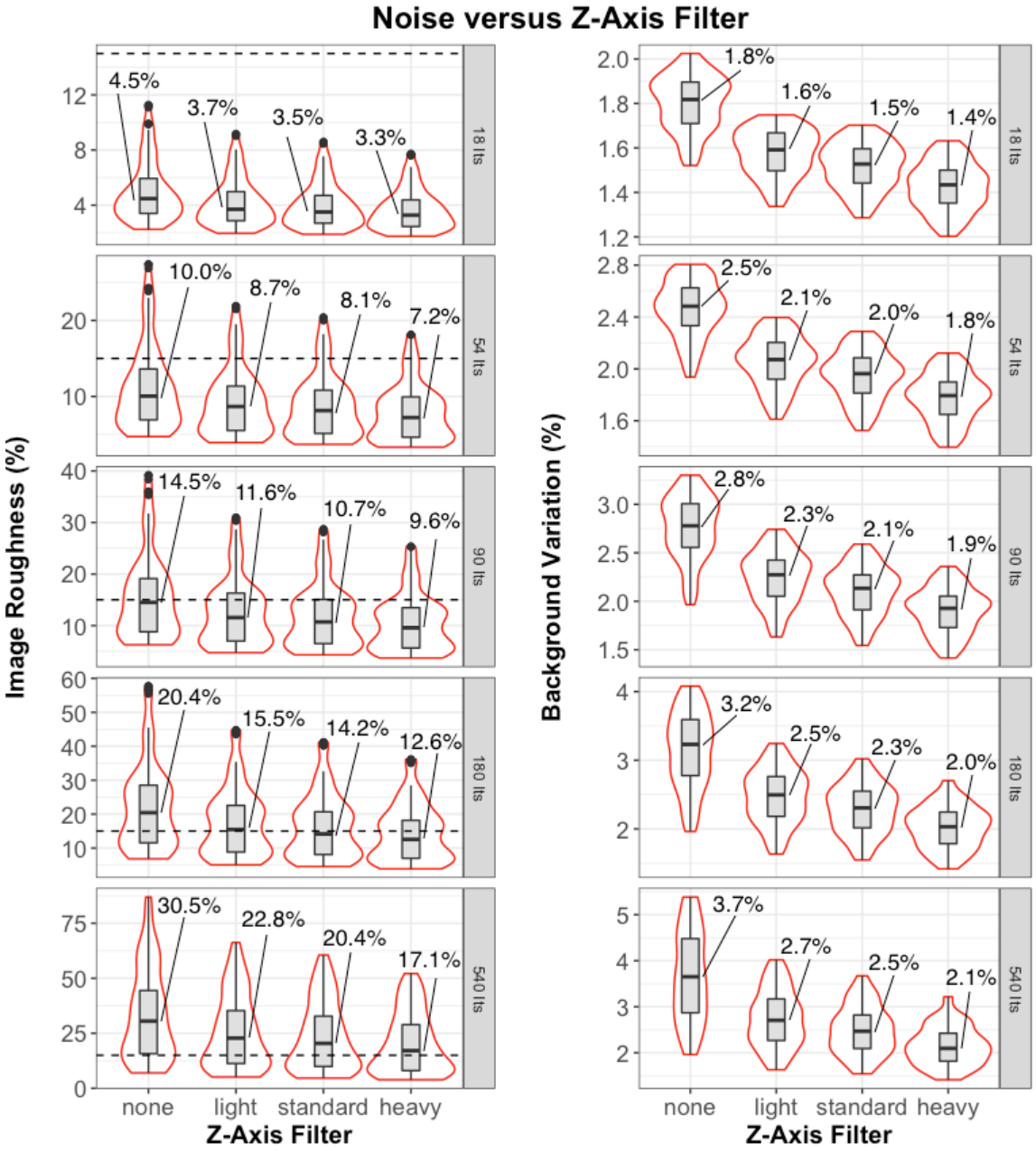


Figure 6.4: Image Roughness and Background Variation versus z-axis filter

Results shown as box and violin plots. Whiskers in box-plots represent 1.5 inter-quartile ranges and dots represent outliers. Each row represents a different number of effective iterations, with different y-axis limits. Dashed lines on Image Roughness plot represents 15% COV limit suggested by guidelines. Annotations show median result for each z-filter setting.

Figure 6.4 demonstrates that as the z-axis weighting increases, both the median and maximum results for Image Roughness decrease. As with increasing Gaussian filter FWHMs, this is as one would expect, as filtering is known to reduce image noise [44] (discussed in Chapter 1). The ‘light’, ‘standard’ and ‘heavy’ median Image Roughness results are reasonably similar to each other at lower iterations but begin to diverge when greater numbers of iterations are applied; for example, all three filter settings have median results within 12% of each other when 18 effective iterations are applied, which rises to 33% at 540 effective iterations. With respect to the 15% COV limit specified by the guidance, all four median Image Roughness results fall below this limit when up to 90 effective iterations are used. When 180 effective iterations are applied, the ’none’ and ‘light’ filter median values exceed 15%, and when 540 effective filters are applied, all median values exceed 15%.

Figure 6.4 also demonstrates that Background Variation results decrease as the z-axis filter weight increases. As with Image Roughness, the ‘light’, ‘standard’ and ‘heavy’ median results are reasonably similar to each other at lower iterations but begin to diverge when greater numbers of iterations are applied; for example, all three filter settings have median results within 14% of each other when 18 effective iterations are applied, which rises to 29% at 540 effective iterations.

The Friedman test returned a p-value of < 0.001 for both Image Roughness and Background Variation; pairwise Wilcoxon signed-ranks (Bonferroni) tests were therefore performed on each possible pair of z-axis filter weights at each iteration setting. The p-value in each case was < 0.001, indicating each Z-axis filter setting produces statistically significantly different Image Roughness and Background Variation results to the others at the 1% level of significance. However, it appears the differences between each filter setting are small when less than 100 effective iterations are used: these differences may not be clinically significant. Comparing Figures 6.3 and 6.4 demonstrates the Gaussian filter has a more significant effect upon Image Roughness than the z-axis filter, while the z-axis filter has a more significant effect on Background Variation than the Gaussian filter.

6.4.1.4 Effects of Matrix Size

Figure 6.5 illustrates the effects of matrix size on both Image Roughness and Background Variation for the full result set. Results are presented as box and violin plots and each row represents a different number of effective iterations. Different y-axis limits are used for each row to better illustrate the effects of each matrix setting. Each box plot is annotated with the median value.

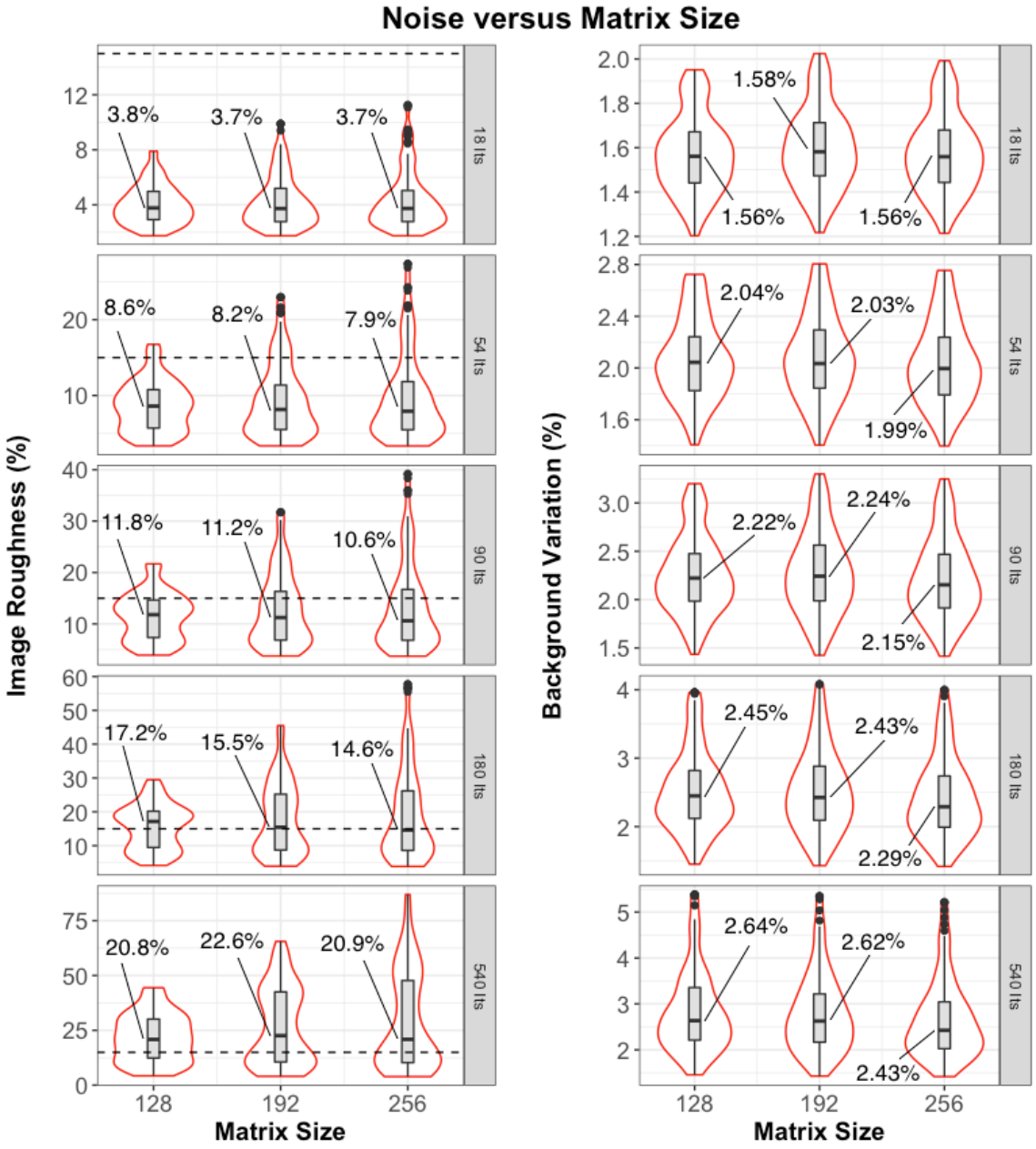


Figure 6.5: Image Roughness and Background Variation versus matrix size

Results shown as box and violin plots. Whiskers in box-plots represent 1.5 inter-quartile ranges and dots represent outliers. Each row represents a different number of effective iterations, with different y-axis limits. Dashed lines on Image Roughness plot represents 15% COV limit suggested by guidelines. Annotations show median result for each matrix setting.

As discussed in Chapter 1, the use of larger matrix sizes (and hence smaller voxels) reduces the number of detected events that contribute to each voxel in the reconstructed image; the statistical noise in each voxel therefore increases as the matrix size increases. One would therefore expect Figure 6.5 to demonstrate that Image Roughness increases as the matrix size increases. Whilst this is true for the outlying results, the median values do not follow the expected pattern: when 180 effective iterations or less are applied, the largest matrix produces the lowest median results and the smallest matrix produces the greatest

median results. With respect to the 15% COV limit specified by the guidance, the median Image Roughness results for all three matrix sizes fall below this limit when up to 90 effective iterations are applied.

Figure 6.5 also demonstrates that the largest matrix size produces the lowest Background Variation results; however, the effects of matrix size upon Background Variation are minimal (maximum discrepancy between median values is only 8% when 540 effective iterations are applied). One would not expect the matrix size to have a significant impact upon region-to-region noise measurements, as the mean value of each ROI is unlikely to significantly change as a result of the change in voxel size.

The Friedman test returned a p-value of < 0.001 for both Image Roughness and Background Variation; pairwise Wilcoxon signed-ranks (Bonferroni) tests were therefore performed on each possible pair of matrix settings at each iteration setting. Only two of the 15 pairs of Image Roughness results demonstrated statistically significant differences (at the 5% level of significance):

- 128 and 192 matrices when 180 effective iterations applied
- 128 and 192 matrices when 540 effective iterations applied

Conversely, only four of the 15 pairs of Background Variation results did not demonstrate statistically significant differences (at the 5% level of significance):

- 128 and 256 matrices when 18 effective iterations applied
- 128 and 192 matrices when 54 effective iterations applied
- 128 and 192 matrices when 90 effective iterations applied
- 128 and 192 matrices when 180 effective iterations applied

Any clinically significant differences as a result of altering the matrix size are likely to result only from outliers at higher iterations, i.e. when no filtering is applied to control the noise. However, the surprising Image Roughness results in Figure 6.5 are investigated in more detail later in this Results section.

6.4.1.5 Effects of TOF and PSF

TOF and PSF reconstructions are known to produce different noise patterns, as discussed in Chapters 1 and 2: TOF produces a random noise pattern, similar to that of HD, while PSF produces a more correlated appearance in background activity [19], [86], [134], [135], [140]. A brief qualitative comparison between the four reconstruction methods is shown in Figure 6.6, while the full quantitative comparison is shown in Figure 6.7. Figure 6.6 compares the central phantom slice from all four reconstruction methods. The remaining reconstruction parameters were kept constant in each case, and match the GEMS suggested settings for

clinical reconstructions (54 iterations, 4mm FWHM Gaussian filter, 192 matrix and the Standard z-axis filter). Image Roughness and Background Variation results for each reconstruction are also included.

	
HD Reconstruction IR = 11.45%; BV = 2.10%	PSF Reconstruction IR = 7.84%; BV = 2.22%
	
TOF Reconstruction IR = 12.30%; BV = 1.91%	PSF+TOF Reconstruction IR = 8.82%; BV = 1.95%

Figure 6.6: Qualitative and quantitative comparisons of phantom images using different reconstruction methods

All four reconstructions used 54 effective iterations, 4mm FWHM Gaussian filter, 192 matrix and Standard z-axis filter (as per GEMS suggestion). Image Roughness (IR) and Background Variation (BV) results shown for all images.

Figure 6.6 demonstrates that, as expected, both PSF reconstructions have smoother, more correlated appearances than the non-PSF reconstructions, which appear to demonstrate a more random noise pattern. This is confirmed by the Image Roughness results which, as expected, are lower for the PSF reconstructions than the non-PSF reconstructions [19], [86], [134], [135], [140]. The TOF-only reconstruction has a greater Image Roughness than the HD reconstruction. As discussed in Chapter 2, TOF is known to reach convergence after fewer iterations than non-TOF reconstructions. This means that when TOF and non-TOF reconstructions using the same number of iterations are compared, the TOF reconstruction may demonstrate greater voxel-to-voxel noise. Furthermore, the phantom size is unlikely to be large enough to fully demonstrate the advantages of using TOF. Studies have shown only minimal TOF noise improvements for phantoms with diameters of approximately 18cm [19], [133]; the phantom used in this chapter has a 20cm diameter.

Figure 6.7 illustrates the effects of TOF and PSF on both Image Roughness and Background Variation for the full result set. Results are presented as box and violin plots and each row represents a different number of effective iterations. Different y-axis limits are used for each row to better illustrate the effects of each reconstruction method. Each box plot is annotated with the median value.

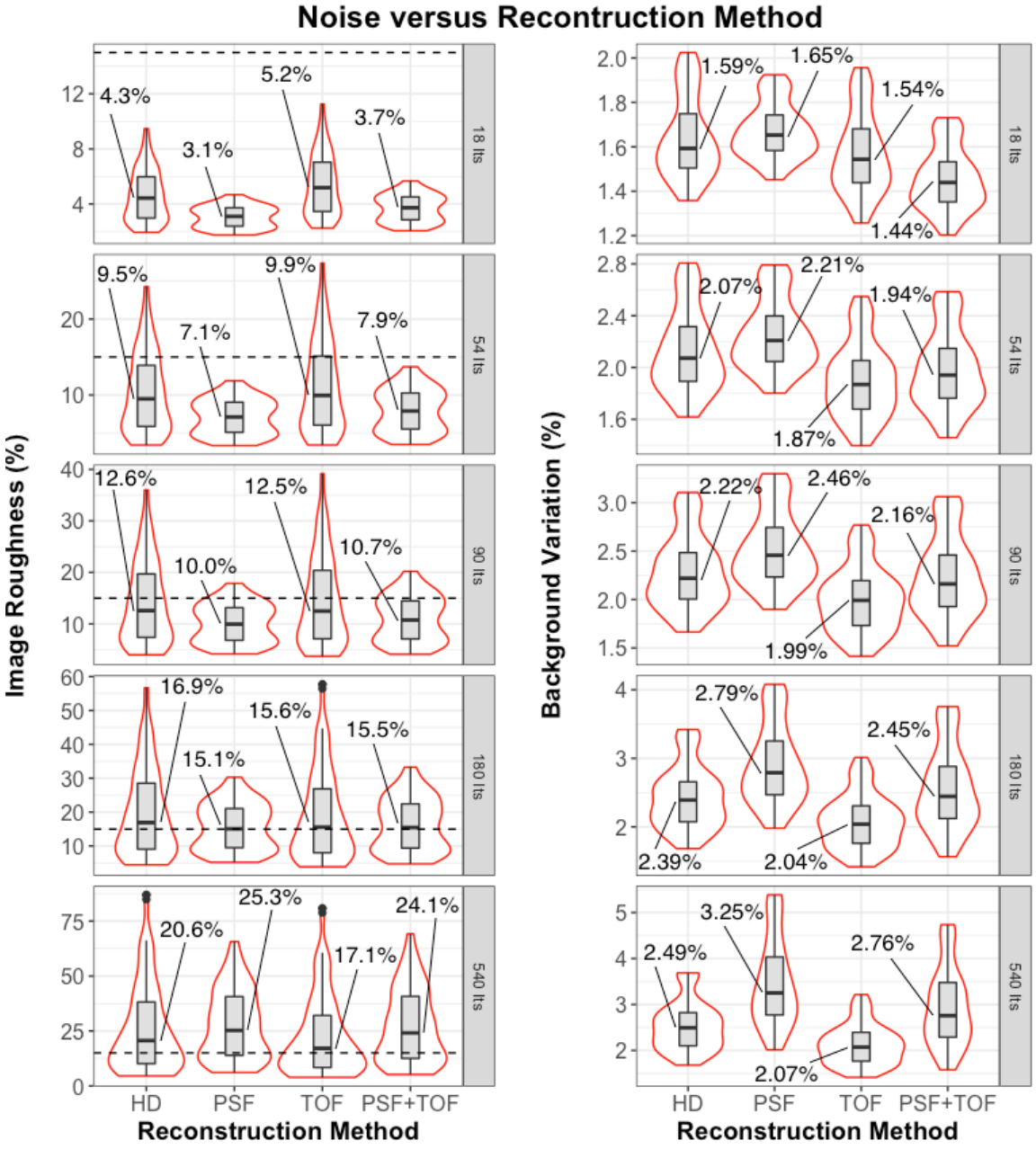


Figure 6.7: Image Roughness and Background Variation versus reconstruction method

Results shown as box and violin plots. Whiskers in box-plots represent 1.5 inter-quartile ranges and dots represent outliers. Each row represents a different number of effective iterations, with different y-axis limits. Dashed lines on Image Roughness plot represents 15% COV limit suggested by guidelines. Annotations show median result for each reconstruction method.

Figure 6.7 demonstrates that at lower iterations (less than 100), PSF clearly reduces Image Roughness relative to HD in terms of both median and maximum values and produces much more consistent results than HD or TOF, as expected. Conversely, the application of TOF increases both median and maximum results relative to HD when up to 54 iterations are applied, also as expected. However, as the number of effective iterations increases, the relative effects of TOF and PSF upon Image Roughness change: at 540 iterations, TOF produces the lowest median result whilst PSF produces the greatest median result. With respect to the 15% COV limit specified by the guidance, all Image Roughness results produced using PSF fall below this limit when up to 54 effective iterations are applied. Median results for all four reconstruction methods are below the 15% limit when up to 90 effective iterations are applied. Applying PSF and TOF together produce similar result distributions to PSF-only, with marginally increased median results as a result of TOF's inclusion at lower iterations. As iterations increase, the effect of TOF upon the combined reconstruction appears to become more significant as it reduces the median value: at 540 iterations, PSF+TOF produces a lower median value than PSF-only.

In contrast to the Image Roughness results, Figure 6.7 demonstrates that PSF produces the greatest median Background Variation result regardless of how many iterations are applied. This is not unexpected; increased inter-voxel correlations caused by PSF may lead to greater differences between ROI mean values, and hence increased Background Variation [19], [92]. TOF produces the lowest median Background Variation when at least 54 iterations are applied. As TOF effectively shortens LORs, noise correlations throughout the FOV are reduced, so this is also as expected [192]. Applying PSF and TOF together largely produces median results between the PSF-only and TOF-only median results. At higher iterations, the PSF+TOF median Background Variation exceed that of HD.

The Friedman test returned a p-value of < 0.001 for both Image Roughness and Background Variation; pairwise Wilcoxon signed-ranks (Bonferroni) tests were therefore performed on each possible pair of reconstruction methods at each iteration setting. All Background Variation pairs demonstrated statistically significant differences, while only one of the 30 Image Roughness result pairs did not demonstrate statistically significant differences (at the 5% level of significance):

- HD and TOF when 90 effective iterations applied

Noise differences between the four reconstruction methods may be clinically significant, particularly at the outlying results where little or no filtering is applied. Although the differences between median results are small, particularly at lower iterations, changes in the appearance of noise may affect an observer's ability to differentiate between lesion and background activity.

6.4.2 Unexpected Relationship between Noise, Matrix Size and Gaussian Filter Width

Conventional wisdom dictates that, when all other reconstruction parameters are kept constant, noise in the 256 matrix images should exceed that of the 192 matrix, which should in turn exceed that of the 128 matrix, because smaller voxels have increased statistical noise (as discussed in Chapter 1). However, the previous section demonstrated that Image Roughness appeared to decrease as the matrix size increased. These results were therefore analysed further to determine if there were any confounding effects caused by combinations of reconstruction parameters.

Figure 6.8 illustrates the relationship between matrix size and Gaussian filter FWHM for Image Roughness. Each row in Figure 6.8 represents a different combination of the remaining reconstruction parameters (reconstruction method, z-axis filter weight and effective iterations). There are 80 different combinations of reconstruction method, z-axis filter and effective iterations in the dataset used for this chapter. All 80 combinations were produced; while only four of these plots (selected at random) are presented in this chapter for brevity, all 80 plots demonstrated the same pattern of results shown in Figure 6.8. The dotted line on the plots indicates the 4mm Gaussian filter width, which is suggested by GEMS for clinical imaging.

At low filter widths (up to approximately 2mm FWHM), Image Roughness results are as one would expect: the 256 matrix produces the greatest Image Roughness whilst the 128 matrix produces the lowest Image Roughness. As the filter width increases, Image Roughness of the larger matrices begin to reduce at a lower filter FWHM than the smaller matrices:

- 256 matrix requires FWHM > 1mm to reduce Image Roughness
- 192 matrix requires FWHM ≈ 2mm to reduce Image Roughness
- 128 matrix requires FWHM ≈ 3mm to reduce Image Roughness

One would expect this to be the case; wider Gaussian filter widths are required to impact the filtered values of larger voxels. One would also expect the smallest voxels to continue to produce the greatest noise levels regardless of the filter applied, until such point that the filter width is large enough to produce similar results for all matrix sizes. However, as the filter width increases beyond approximately 3mm, the relative Image Roughness pattern of the three matrix sizes begins to follow an unexpected pattern:

- 256 matrix Image Roughness falls below 192 matrix at FWHM ≈ 3mm
- 256 matrix Image Roughness falls below 128 matrix at FWHM ≈ 4mm
- 192 matrix Image Roughness falls below 128 matrix at FWHM ≈ 4.5mm

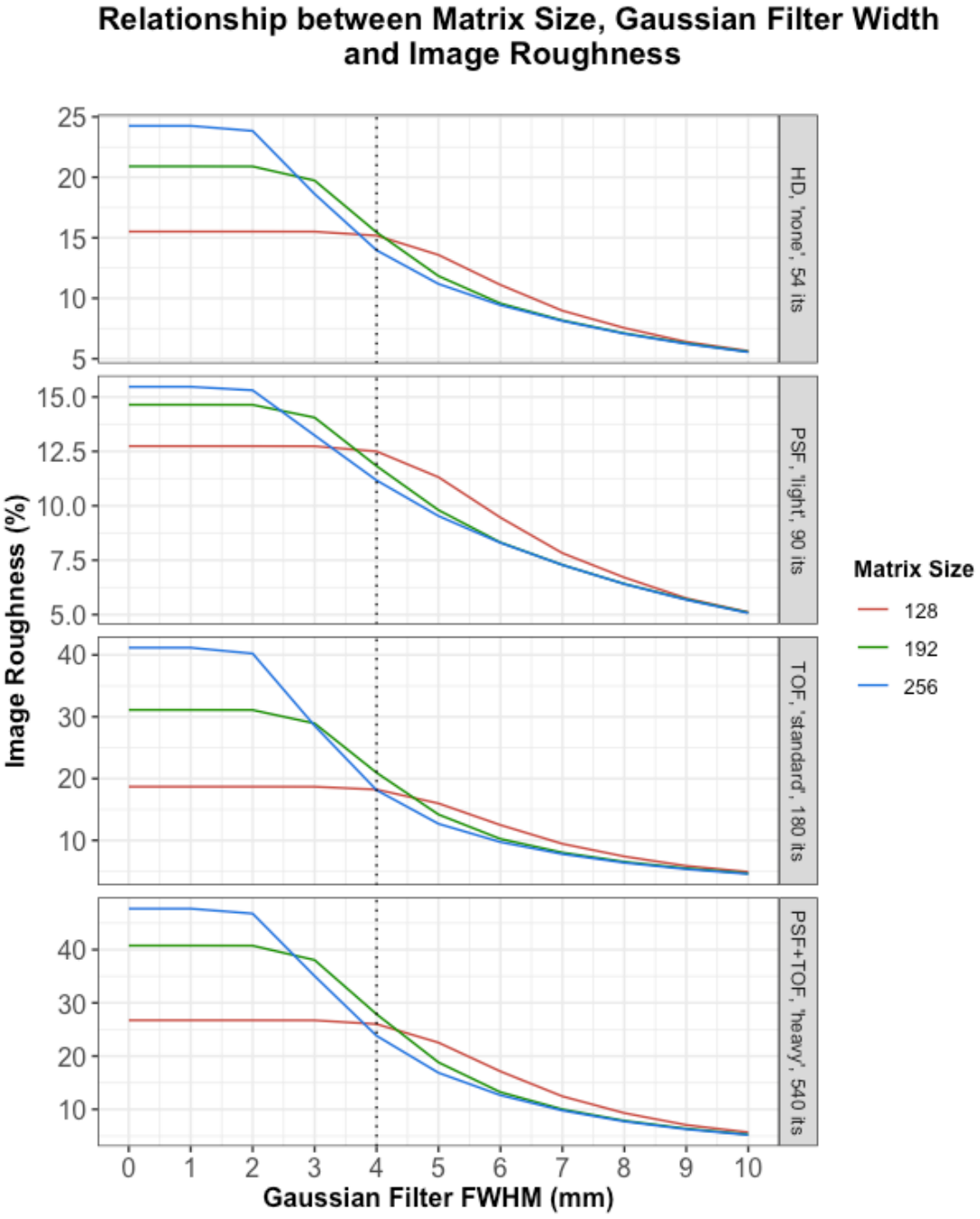


Figure 6.8: Unexpected relationship between Image Roughness, Gaussian filter width and matrix size

Each row represents a different combination of remaining reconstruction parameters. Dotted lines indicate GEMS recommended Gaussian filter width for clinical imaging.

At filter widths greater than ≈ 4.5mm, the expected Image Roughness noise pattern is entirely reversed: the smallest voxels produce the lowest Image Roughness, while the largest voxels produce the greatest Image Roughness. At larger filter widths (≈ 9mm), all three matrix sizes produce very similar results.

As this relationship between Image Roughness, matrix size and filter width was an unexpected finding, an additional analysis was performed on a subset of the reconstruction using Matlab instead of Hermes Hybrid Viewer, in order to rule out any problem with the measurement technique. The Matlab analysis produced the same relationship between noise, matrix size and Gaussian filter width.

6.4.2.1 ^{18}F-FDG Patient Liver Noise Analysis

A small retrospective patient study was performed to verify this unexpected relationship between noise, matrix size and Gaussian filter width wasn't limited to the ^{68}Ge cylindrical phantom experiment. Ten consecutive patients reported to have no liver abnormalities were selected. All patients were fasted for at least 6 hours and received intravenous injections of 400MBq +/- 10%. Imaging was performed 60 minutes post-injection. Patient BMIs ranged from 20.6 to 35.4. Each set of patient data was retro-reconstructed 33 times, as follows:

- Gaussian filter FWHM: varied from 0mm to 10mm, in 1mm increments
- All three matrix sizes: 128, 192, 256
- Remaining reconstruction parameters kept constant: HD reconstruction, 54 effective iterations, and no z-axis filter applied. Transaxial FOV kept constant at 700mm. These reconstruction parameters matched those of the first graph in Figure 6.8.

Image Roughness for each patient liver was analysed by placing three 3cm diameter ROIs on three consecutive transaxial slices of the liver. The ROIs were positioned in visibly uniform areas of the liver, avoiding major blood vessels. This method of liver noise analysis has previously been used in studies by Akamatsu et al [132], [134], [146] and Taniguchi et al [145].

Figure 6.9 shows a transaxial liver slice from the patients with the smallest and largest BMI, as well as their Image Roughness versus Gaussian FWHM plots. All ten patient analyses demonstrated similar findings: the unexpected relationship between noise, matrix size and Gaussian filter width is also observed for patient data. As one would expect, Figure 6.9 also demonstrates the noise in the larger patient is greater than that of the smaller patient, both qualitatively and quantitatively.

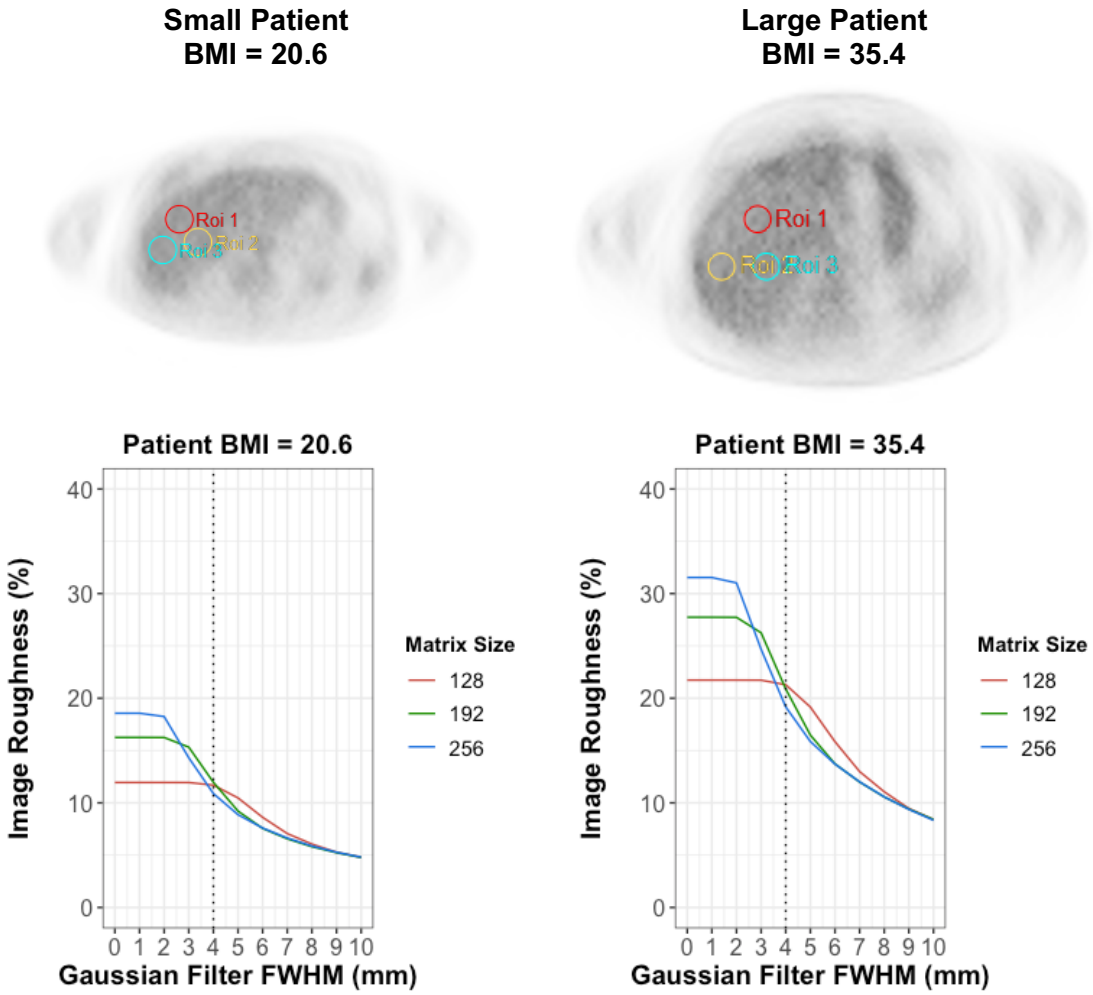


Figure 6.9: Patient liver images and Image Roughness results

Example reconstruction images used 4mm FWHM Gaussian filter and 192 matrix.

6.4.2.2 Correspondence with GEMS Engineers

Dr Charles Stearns, the senior engineer at GEMS, was contacted directly about these unexpected findings (personal correspondence, 1st August 2016). As Dr Stearns had not encountered this phenomenon before, he performed his own simulation of the GEMS Gaussian filter operation for different matrix sizes and filter FWHMs, and produced the same unexpected relationship between noise, matrix size and filter FWHM demonstrated by the work in this chapter.

Dr Stearns explained the Discovery 690's Gaussian filter is implemented by creating a Gaussian curve of the specified FWHM and selecting samples from the curve at intervals that correspond to the voxel widths. The filter is truncated to ± 4σ, where σ is the standard deviation, with a minimum of three points in the kernel, normalised to give a total of 1.0. As a result of this implementation method, the filter is close to [0 1 0] at lower filter widths. As the filter width increases, it operates more like a three-point averaging filter instead of a true

Gaussian. The Gaussian filter is therefore not a true Gaussian. The sampling of the Gaussian curve using the voxel width causes the unexpected relationship between noise, matrix size and filter FWHM.

Dr Stearns simulated the Image Roughness results that would be obtained using an ideal Gaussian filter, shown in Figure 6.10 below. The ideal filter is designed as the integral under the Gaussian for each voxel, instead of simply sampling a point from the Gaussian curve for each voxel. This implementation produced the expected relationship between different matrix sizes; noise in the larger matrices always exceeded that of the smaller matrices until such point that the filter width is large enough to produce similar noise results for all matrices.

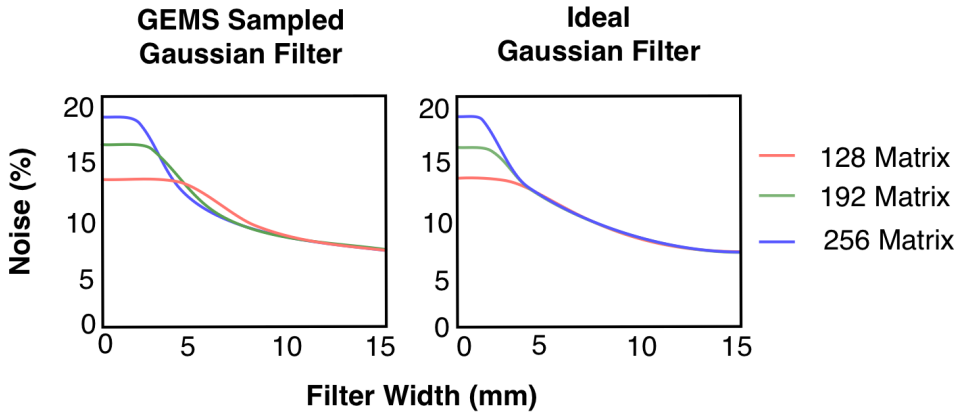


Figure 6.10: GEMS sampled Gaussian filter versus ideal Gaussian filter Adapted from simulations provided by Dr Charles Stearns (personal communications, 1st August 2016).

6.5 Discussion

Although the effects of effective OSEM iterations, filtering, voxel size, PSF and TOF upon PET image noise had been widely assessed in the literature, no single study had examined the effects of all of these parameters in combination. In particular, the combined effect of matrix size and filters on PET image noise had not yet been evaluated fully. Furthermore, there remains no universally agreed approach to clinically relevant noise assessment in PET imaging. This chapter therefore aimed to characterise the combined effects of these reconstruction parameters on both pixel-to-pixel noise (Image Roughness) and region-to-region noise (Background Variation) in a clinically relevant manner and assessed the use of the widely available 20cm diameter uniform ^{68}Ge phantom for this purpose.

PSF was found to improve Image Roughness but had a small detrimental effect upon Background Variation, similar to results from previous studies [19], [86], [134], [135], [140].

This remained the case even when combined with TOF, when at least 54 effective iterations were applied. The clinical implications of this finding are difficult to determine by assessing noise in isolation; the effects of PSF noise characteristics upon lesion detection are examined later in this thesis. For example, human observers may not agree that the correlated background activity produced by PSF improves image quality with respect to lesion detection, even though the Image Roughness metric suggests an improvement. TOF was shown to produce similar, but increased, Image Roughness results to the HD reconstruction when up to 180 effective iterations were applied. While this effect is partly caused by the early convergence of TOF, the 20cm diameter of the uniform phantom is not large enough to fully demonstrate the advantages of using TOF [19], [133].

The assessment of the remaining reconstruction parameters largely agreed with previously published studies. The number of applied effective iterations had a significant effect upon both Image Roughness and Background Variability: both noise metrics increased as the number of effective iterations increased, as expected [43]. Both the Gaussian filter and the z-axis filter were shown to reduce image noise, also as one would expect [44]. Z-axis filtering was shown to have a greater effect on Background Variation than the Gaussian filter, while the Gaussian filter was shown to have a greater effect on Image Roughness than the Z-axis filter. This suggests the Z-axis filter should be given greater consideration in mitigating the detrimental effects of PSF on Background Variation. As discussed in Chapter 2, the effect of Z-axis filtering is not as widely discussed in the literature as filtering in the transaxial plane, as z-axis filtering has not been adopted by all vendors.

One would expect the use of larger matrices, and hence smaller voxels, to result in increased image noise, as discussed in Chapters 1 and 2. However, the effects of matrix size were not as expected; the larger matrix produced the smallest median results for Image Roughness. Only the outlying results, corresponding to reconstructions which used little or no filtering, followed the expected noise pattern as the matrix size increased. An unexpected relationship, previously unknown to even GEMS, was therefore identified after further analysis. At certain filter widths (between approximately 3mm and 9mm FWHM), Image Roughness in the larger voxels exceeds that of the smaller voxels. This may have clinical relevance as GEMS suggest using a 4mm filter with a 192 matrix for clinical image reconstruction. If an imaging centre who initially followed the GEMS suggested reconstruction strategy wished to increase the matrix size from 192 to 256, e.g. with a view to improving the spatial resolution of their images, they would expect the image noise to increase. The work in this chapter has shown that, in fact, the measured noise would decrease. This raises questions as to how the change in matrix size would affect spatial resolution: if using smaller voxels unexpectedly improves the noise, could spatial resolution be unexpectedly worsened? Spatial resolution is investigated over the next two chapters.

All Image Roughness results were compared with the 15% COV limit recommended by European guidance. The results in this chapter demonstrated that the number of effective iterations, the Gaussian filter width and the use of PSF were the dominant parameters in terms of achieving this limit. At 54 effective iterations, as suggested by GEMS for clinical reconstructions, all Image Roughness results were below the 15% limit when PSF was used, even when no filtering was applied. The majority of PSF results at 90 effective iterations were also below this 15% limit (the limit was only breached when a Gaussian filter width of less than 4mm was combined with no z-axis filtering for PSF-only). The GEMS suggested reconstruction parameters for clinical imaging (54 effective iterations, PSF+TOF, 4mm Gaussian, "standard" z-axis filter and 192 matrix) produced 8.8% Image Roughness, below the 15% limit; this fell to 8.4% when a 256 matrix was used instead.

European guidance on ^{18}F-FDG imaging, discussed in Section 2.1, state that matrix sizes and zoom factors should be chosen such that reconstructed voxel sizes should be between 3mm and 4mm in any direction. The reason for this is not explicitly stated but is presumably to achieve acceptable spatial resolution whilst suppressing noise in the reconstructed images. Whole-body imaging on the GEMS Discovery 690 typically requires a 700mm transaxial FOV (the z-axis FOV is fixed at 157mm, with axial sampling fixed at 3.34mm). This means that only the 192 matrix complies with this voxel size recommendation:

- 128 matrix produces voxel sizes 5.47mm x 5.47mm x 3.34mm
- 192 matrix produces voxel sizes 3.65mm x 3.65mm x 3.34mm
- 256 matrix produces voxel sizes 2.73mm x 2.73mm x 3.34mm

The EANM guidelines also state that the maximum Gaussian filter width should be 7mm; however, no recommendations are made in terms of combinations of filters and matrix size. The guidelines may therefore merit review as a result of the unexpected relationship between noise, filters and matrix size, particularly if the motivation for restricting voxel size was to limit image noise rather than reconstruction time. The use of the 256 matrix, and hence sub 3mm voxel sizes, may be more appropriate for whole-body imaging using the GEMS Discovery 690.

The major limitation of this study was the use of a single, relatively small phantom. A larger phantom is required to fully assess the effects of TOF upon image noise; larger phantoms would also provide more realistic representations of patient body habitus for whole-body PET imaging. A larger phantom will be inherently noisier; this will therefore also affect which reconstruction parameters achieve the 15% EANM noise limit. However, it should be noted that the count density of the phantom used in this chapter was conservative when compared to that of ^{18}F-FDG patient livers, as discussed in Chapter 4; statistical noise levels within the phantom were therefore clinically relevant, in spite of its relatively small size.

Furthermore, the results of this phantom may be applicable to other clinical applications, for example paediatric or head and neck imaging. Both Image Roughness and Background Variation are assessed using a larger phantom in Chapter 9 (alongside other image quality metrics). Patient data is also used to assess noise later in this thesis.

6.6 Conclusions

This chapter concludes that, with the exception of matrix size, the effects of the reconstruction parameters upon both Image Roughness and Background Variation were as expected, and consistent with the literature.

This chapter further concludes that an unexpected relationship exists between matrix size, Gaussian filter width and noise measurements on the GEMS Discovery 690 PET-CT imaging system. Although the effect is small, it may cause confusion when assessing the effects of increasing matrix sizes; for example, when attempting to improve spatial resolution of reconstructed images. Furthermore, this effect may be of clinical interest as it applies to the GEMS suggested Gaussian filter width (4mm). The reasons for this phenomenon, previously unknown to even the GEMS engineers, were confirmed by GEMS as being a result of the sampled Gaussian filter implementation.

Furthermore, this chapter concludes that the EANM guidelines for voxel sizes may merit review. To achieve the recommended voxel size of between 3mm and 4mm on the GEMS Discovery 690 when the full 700mm FOV is used requires the use of the 192 matrix, which can produce greater noise than the 256 matrix when combined with specific Gaussian filter widths. The effects of the GEMS Gaussian filter implementation discovered in this chapter may therefore undermine the intentions of this voxel size guideline.

Finally, this chapter concludes that the 20cm diameter cylindrical phantom is not large enough to fully demonstrate the effects of TOF when assessing image noise; however, the results remain applicable to imaging of smaller structures (e.g. head and neck or paediatrics). A larger phantom is used in the latter chapters of this thesis to assess noise in a manner which is more clinically relevant to liver imaging.

Further work should involve assessing the relationship between matrix size, filtering and image noise measurements using PET systems from other vendors. This would determine if the unexpected relationship observed in this chapter is unique to GEMS or more widespread throughout the PET imaging community.

It should also be noted the analysis in this chapter focussed on five image slices at the centre of the axial FOV, where sensitivity is at its maximum. Further assessment of the z-axis filter should involve assessing its effects within the low-sensitivity, higher-noise overlap area between bed positions.

An abstract based on some of this chapter's work was accepted for a poster presentation at the 2016 Institute of Electrical and Electronic Engineers (IEEE) Medical Imaging Conference.

Chapter 7 : Development of Methodology for Clinical Spatial Resolution Assessment

Spatial resolution is a measure of an imaging system's ability to accurately distinguish between two close together objects and observe their details [69]. The FWHM of the PSF is generally taken as a measure of spatial resolution [32]. Spatial resolution tests are performed to NEMA standards during acceptance testing of new PET-CT systems (and are typically performed on an annual basis thereafter): point sources are imaged in air at different positions within the FOV and reconstructed using FBP. The point source response FWHM is then measured in the x, y and z directions [177]. This standardised approach to spatial resolution measurement allows the comparison of a system's performance against manufacturer's specifications and other different imaging systems, and has been used by several studies to assess spatial resolution [19], [111], [155], [193], [194]. However, NEMA methodology does not provide a clinically relevant assessment of spatial resolution, for reasons outlined in this chapter.

Measuring spatial resolution in a clinically relevant manner is challenging. Measurements should be able to incorporate scatter material and differing source-to-background ratios, use iterative reconstruction parameters to reconstruct the images, and use a clinically realistic FOV. As the capillary tubes used to create NEMA point sources are difficult to place inside background activity, many studies have used alternative methods to assess clinical spatial resolution. However, a review of the literature demonstrated there is no established method for assessing spatial resolution in a clinically relevant manner.

The primary aim of this thesis is to develop a generic methodology to assess and optimise PET image reconstruction whilst considering clinical context, whilst the secondary aim is to

use this methodology to optimise reconstruction on the GEMS Discovery 690 for the detection of small liver lesions. As the assessment of clinical spatial resolution is particularly challenging, two chapters in this thesis have been dedicated to it:

- This chapter investigates methodologies for assessing spatial resolution in a clinically relevant manner, and concludes by recommending the use of a particular methodology
- The next chapter (Chapter 8) uses the methodology developed in this chapter to assess the effects of reconstruction parameters upon spatial resolution when a GEMS Discovery 690 is used for clinical imaging.

7.1 Introduction

The current NEMA methodology does not provide a clinically relevant assessment of spatial resolution for the following reasons:

- Point sources (≤ 1mm diameter) are in air without scatter material or background activity, which is clinically unrealistic in terms of activity distribution, attenuation and scatter. Spatial resolution can be contrast dependent so measured spatial resolution may also vary as the source-to-background ratio varies [94].
- FBP reconstructions used for point source images are not used clinically; the effects of clinical reconstruction parameters are therefore not assessed by the NEMA method. OSEM reconstruction methods should not be used to reconstruct point sources in air because they are known to produce very small, clinically unrealistic results (OSEM convergence is affected by the activity distribution, count rate and spatial frequencies within the FOV) [36], [195], [196]. Spatial resolution in FBP images is also independent of any activity distribution in the background, while iterative reconstruction spatial resolution is dependent upon other objects within the FOV [94].
- The NEMA method stipulates that voxel sizes should be no more than one third of the expected spatial resolution in all three dimensions [70]. The GEMS Discovery 690 protocol for NEMA spatial resolution measurement uses a 200mm transaxial FOV with a 256 matrix to achieve voxel dimensions of 0.78mm in the x and y directions (the z-direction voxel dimensions, however, remain fixed at 3.34mm). A 200mm FOV is not practical for adult whole-body imaging, so these voxel sizes are highly unlikely to be used for clinical liver imaging.

Several studies have attempted to address the challenges of measuring spatial resolution in a clinically relevant manner. Some studies have used long-lived ^{22}Na point sources encased in small volumes of Lucite (typically ≈ 1cm^3), without background activity or scatter material [33], [49], [86], [197]. The point source's FWHM is then measured in the x, y and z directions. This type of source has been shown to approximate an ^{18}F point source in water

[49], so the point source is not strictly 'in air'. This may improve the ^{22}Na point source's clinical relevance compared to ^{18}F capillary tubes in air; however, scatter material is limited to ≈ 1cm^3, and therefore remains clinically unrealistic. One study placed the ^{22}Na source inside a water-filled phantom to provide scatter material [198]. To the author's knowledge, no study has placed a ^{22}Na point source within ^{18}F background activity.

Line sources have also been used in the literature to assess spatial resolution. The older NEMA 1994 standard specified that spatial resolution should be measured using a stainless steel line source phantom, with the line sources suspended in air [69], [171]. However, later NEMA standards moved on to using point sources, as slight misalignments of the line sources within the image matrix could have degrading results [195]. A further disadvantage of using line sources is that FWHM can only be measured in the x and y directions. ^{18}F line sources surrounded by air [199]–[201], water [200] or background radioactivity [49], [142], [195] have all been used to assess spatial resolution. Another study used a long-lived ^{68}Ge line source, which was imaged in both air and within a water-filled Jaszczak phantom [198].

Perturbation methods for measuring spatial resolution of iterative reconstruction SPECT images have been used in the literature [202]–[204]. This method is used to minimise the distortions that would be introduced by iterative techniques if point/line sources were imaged in air. The term 'perturbation' refers to the fact that the line/point sources are 'perturbed' by the presence of background activity. Placing line sources within a background activity instead of placing them in air mitigates the context-dependent problems associated with iterative reconstruction. Such methods typically involve producing 'unperturbed' data sets (line/point source only) and adding background activity either through simulations [202] or by manipulating the projection data of separate source and background acquisitions prior to reconstruction [203], [204]. This creates a 'perturbed' data set (sources in the presence of background activity). Subtracting the background activity from the 'perturbed' data set produces a background-corrected, source-only data set which can then be used for analysis. More recently, a study by Brown et al [205] used a modified version of the perturbation technique for SPECT imaging. A line source phantom containing ^{99m}Tc was scanned twice. For the first acquisition, the phantom contained 'background' activity that surrounded empty line sources. The line sources were then filled in situ and the phantom was scanned again to produce the 'perturbed' data set. The reconstructed images of the first scan (surrounding activity only) were then subtracted from the second scan ('perturbed' line sources) to produce a source-only data set, which was then used for analysis.

To the author's knowledge, no study has applied the perturbation method to PET data. Each of the three papers mentioned earlier in this chapter that placed line sources within surrounding radioactivity used the term 'background activity' to describe the surrounding

activity; however, this may not be technically correct. The 'background' activity in each of these papers appears to surround the line sources, but is not co-located with the line sources, and is therefore not true background activity (the line source is not superimposed on top of a background activity). Analysis of planar nuclear medicine images of such a phantom may be correct to use the term 'background' activity, depending on the phantom's orientation; however, both SPECT and PET imaging produce cross-sectional images. 'Surrounding activity' is therefore a more accurate term to describe these line source phantoms when cross-sectional images are involved. Two of these papers briefly discuss performing 'background subtraction', as follows:

- "…resolutions were determined taking the background of the warm cylinder into account" – de Jong et al [195]
- "For the profiles through the warm background images, we subtract the baseline background activity." – Alessio et al [49]

Neither paper discusses how this 'background subtraction' was performed or describes any additional acquisitions of the phantom without line source activity that could be subtracted from the perturbed data.

Analysis methods in the literature for both point and line sources invariably involve drawing line profiles through the point or line sources, fitting the profiles to Gaussian functions, and measuring the FWHM of the fitted Gaussian curves [53], [206]–[208]. Figure 7.1 compares the theoretical effects of 'simplistic background subtraction' (right-hand column) with subtraction of surrounding activity (left-hand column) upon the measured FWHM when a line source phantom is used ('perturbation method').

Figure 7.1 demonstrates that in a zero-noise, high spatial resolution scenario, the perturbation method should result in a line-source-only image in which the line source peak and FWHM are unaffected by the subtraction of the surrounding activity. However, when simplistic background subtraction is used, the maximum point of the fitted Gaussian is reduced by an amount that is dependent on the 'background' activity being subtracted. This in turn reduces the 'half maximum' point in the curve at which the FWHM is measured. Simplistic background subtraction may therefore cause the FWHM to be measured at a position that is closer to the fitted Gaussian's peak, which in turn may produce a narrower FWHM result than the perturbation method for the same line source acquisition. The difference between the perturbation method and simplistic background subtraction results will be more apparent when low line-to-background ratios are used. However, it is not clear if any FWHM discrepancies between the two methods would be affected by alterations in the reconstruction algorithm.

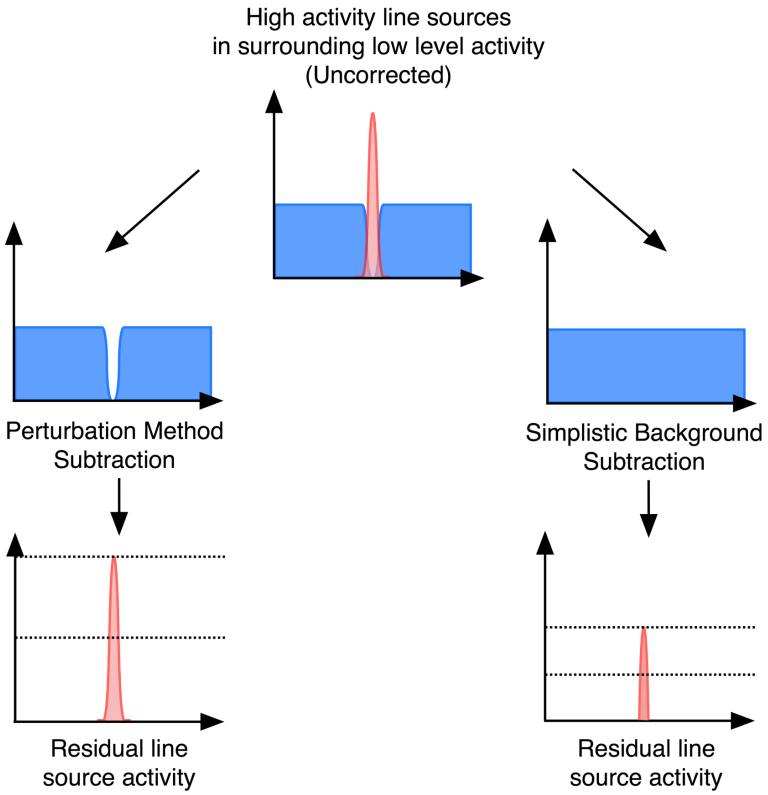


Figure 7.1: Illustration of perturbation and simplistic background subtractions

Nyquist theory states that as an absolute minimum for sampling, voxel sizes must be no greater than half the size of the smallest object to be detected. Clinical liver imaging on the GEMS Discovery 690 typically requires the largest possible (700mm) transaxial FOV (although this may be reduced for smaller patients). The smallest possible voxel dimension in the transaxial plane is therefore 2.73mm (when the 256 matrix is used). This implies that objects less than 5.46mm will be under-sampled. As voxel sizes increase, the size of the smallest detectable object also increases, leading to a loss of high frequency information in the reconstructed images (previously discussed in Chapter 1). Voxel size becomes the major limiting factor for spatial resolution once they are increased above a certain threshold. To the author's knowledge, nobody has assessed the effects of increasing voxel sizes from the minimum to maximum possible sizes in order to determine the threshold where voxel size itself becomes the main spatial resolution limitation. It is therefore unclear from the literature where this voxel size threshold is for the GEMS Discovery 690.

7.2 Aims

The literature review demonstrated there is no established method of measuring PET spatial resolution in a clinically relevant manner. This chapter therefore aims to determine an appropriate phantom measurement technique to allow assessments of the relative differences in clinical spatial resolution resulting from the use of different iterative

reconstruction parameters. It should be noted that this chapter does not aim to produce a definitive numerical value to describe the clinical spatial resolution of PET images; the following chapter will use the methodology developed in this chapter to assess clinical spatial resolution.

Firstly, a ^{22}Na point source and an ^{18}F line source will be used to assess the effects of voxel size on spatial resolution measurements. Specifically, this experiment aims to demonstrate the point at which voxel size becomes a limiting factor on FWHM measurement using each source. Results will be discussed with respect to voxel sizes used for clinical liver imaging and used to determine which matrix sizes are appropriate for clinical spatial resolution measurement.

The remainder of this chapter will then assess two different phantoms, and three different methods, for spatial resolution measurement.

A ^{22}Na point source surrounded by ^{18}F activity will be assessed:

1. The effects of using ^{18}F correction factors upon the ^{22}Na point source will be assessed, to justify the use of two different radioisotopes within a single phantom.
2. The effects of increasing the activity concentration of surrounding ^{18}F activity upon the measured FWHM will be assessed, to determine if clinically relevant background activities have an effect upon spatial resolution measurement.
3. A comparison between 'uncorrected' and 'simplistic' corrections (see Figure 7.1) for surrounding activity will be made, to assess if the correction is necessary for spatial resolution measurement.
    - The 'perturbation' method will not be assessed for this phantom, for reasons described in the following section.
4. The ability of the phantom to demonstrate the effects of altering reconstruction parameters will be assessed briefly; FWHM measurements from high, medium and low resolution reconstructions will be compared.

An ^{18}F line source phantom with surrounding ^{18}F activity will then be assessed:

1. A comparison between 'uncorrected', 'perturbation' and 'simplistic' corrections for surrounding activity will be made, to assess if corrections are necessary for spatial resolution measurement.
2. The ability of the phantom to demonstrate the effects of altering reconstruction parameters will be assessed briefly; FWHM measurements from high, medium and low resolution reconstructions will be compared.

3. Two different line-to-background ratios will be compared to determine if the use of a single ratio could be sufficient to assess the effects of reconstruction parameters on spatial resolution.

Finally, the results from both phantoms and the different corrections for surrounding activity will be compared. A recommended methodology for spatial resolution measurement will then be made. This method will then be used to conduct an assessment of the effects of reconstruction parameters upon spatial resolution, which forms the next chapter of this thesis.

7.3 Materials and Methods

All acquisitions and reconstructions were performed on the GEMS Discovery 690 PET-CT system. A detailed description of this system is included in Chapter 3.

7.3.1 Phantoms

Two phantoms were assessed: a line source phantom and a point source phantom.

7.3.1.1 Point Source Phantom

As demonstrated in the literature review, no study has placed a long-lived point source within surrounding ^{18}F activity. A long-lived point source phantom was therefore created by placing a ^{22}Na point source inside the NEMA body phantom, allowing the point source to be surrounded by plain water or radioactivity. This phantom is described in more detail in Section 3.2.3.1. The ^{22}Na source activity was 0.525MBq at the time of scanning (activity concentration of the 0.25mm diameter sphere was therefore approximately 64GBq/ml).

The perturbation method was not attempted with the point source phantom: this would require removal of the phantom's lid to substitute the ^{22}Na for a space-saver of equivalent dimensions and positioning (thus preserving the 'surrounding activity' nature of the phantom). The phantom would have to be removed from the scan room for the adjustment and then repositioned on the scanner, which could introduce alignment issues. Removing the phantom's lid whilst it contained radioactivity would also have radiation protection implications and could cause the loss of some of the radioactivity. It was therefore not practical to perform the perturbation method on this particular phantom; only the uncorrected data and the simplistic background correction data were assessed.

7.3.1.2 Line Source Phantom

The NEMA 1994 line source phantom contains two fillable line sources (diameter 1.2mm) within a fillable cylindrical cavity. One source is positioned at the phantom centre, while the other is positioned 75mm radially (near the phantom periphery). This phantom is described in more detail in Section 3.2.3.2. Both line sources were filled for this chapter's experiments.

7.3.2 Image Acquisition Protocols

7.3.2.1 Point Source Phantom

The phantom was filled with plain water and positioned with the point source at the FOV centre. The phantom was acquired twice without moving the phantom: once using ^{22}Na correction settings and once using ^{18}F correction settings. A single acquisition frame and 4-minute bed-times were used for both scans. ^{18}F was then added incrementally to the phantom to produce five different source-to-background ratios, detailed in Table 7.1. Approximate source-to-background ratios ranged from 7,500,000:1 to 31,000,000:1. For clinical relevance, the final two surrounding activity concentrations were chosen to be similar to that found within healthy livers of patients undergoing 400MBq ^{18}F-FDG scans, while first three activity concentrations were chosen to provide a range of source-to-background ratios.

Surrounding Activity Concentration	Point Source Activity Concentration	Ratio (approx)	Point Image Maximum Activity Concentration	SUVmax
2,050Bq/ml	64GBq/ml	31M:1	4,261,905Bq/ml	2,236
3,969Bq/ml	64GBq/ml	16M:1	3,379,326Bq/ml	923
5,462Bq/ml	64GBq/ml	12M:1	4,304,010Bq/ml	845
7,237Bq/ml	64GBq/ml	9M:1	4,365,974Bq/ml	643
8,548Bq/ml	64GBq/ml	7.5M:1	4,191,521Bq/ml	524

Table 7.1: Point source phantom activity concentrations

Resultant SUV_{max} values when GEMS' suggested reconstruction parameters are applied are also included.

Each acquisition was performed using ^{18}F correction factors, a single acquisition frame and 4-minute bed-times. It should be noted that while the source-to-background ratios described here appear unrealistically high with respect to clinical imaging, it was anticipated that, following convolution with the imaging system's PSF and post-reconstruction filtering, the reconstructed images would produce much lower effective source-to-background ratios. Preliminary reconstructions using GEMS' suggested clinical parameters demonstrated the point source SUV_{max} values were much lower than the true ratios suggested but were still greater than one would expect to find clinically: this was a result of the high point source

activity concentration. Maximum activity concentrations and SUV_{max} measured from the point source images are also included in Table 7.1.

7.3.2.2 Line Source Phantom

The phantom's main chamber was filled with an activity concentration similar to that found in 400MBq ^{18}F-FDG patient liver images. A first scan was performed with empty line sources to produce an image of the surrounding activity only. A second scan was then performed after filling the line sources in situ, producing a perturbed line source image (the phantom was not moved between scans and identical axial FOV coverage was used for both scans). This set-up was repeated on two separate days with two different line-to-background ratios. A single acquisition frame and 4-minute bed-times were used for all scans. Table 7.2 details the activity concentrations used: ratios were approximately 200:1 and 100:1 respectively. Preliminary reconstructions using GEMS' suggested clinical reconstruction parameters verified the line source SUV_{max} values were similar to what one would expect to observe in a liver lesion. Maximum activity concentrations and SUV_{max} measured from the line source images are also included in Table 7.2.

Surrounding Activity Concentration	Line Source Activity Concentration	Ratio	Line Image Maximum Activity Concentration	SUVmax
8,316Bq/ml	1,609,575Bq/ml	194:1	53,124Bq/ml	7.23
8,163Bq/ml	850,996Bq/ml	104:1	30,249Bq/ml	4.22

Table 7.2: Line source phantom activity concentrations

Resultant SUV_{max} values when GEMS' suggested reconstruction parameters are applied are also included.

7.3.3 Image Reconstruction Protocols

The ^{22}Na plain-water phantom and the 200:1 ^{18}F line source phantom were used to determine the voxel size threshold for spatial resolution limitation. High-resolution reconstructions were performed (PSF with no filtering) with voxel sizes ranging from 0.25mm (minimum possible) up to 5.47mm (maximum possible) in 0.25mm increments. Voxel sizes were achieved by varying matrix size and transaxial FOV. Effective iterations were varied from 18 up to 540, as spatial resolution is known to improve with increased iterations [53].

All remaining reconstructions in this chapter were performed using a 700mm FOV and a 256 matrix (voxel dimensions were 2.73mm x 2.73mm x 3.34mm): the reasons for this will be justified in Section 7.4.1.

The ^{22}Na plain-water phantom acquisitions used to compare the use of ^{22}Na and ^{18}F correction factors were reconstructed using GEMS' suggested clinical reconstruction parameters (PSF+TOF, 54 effective iterations, 4mm Gaussian filter, Standard z-axis filter).

A limited set of reconstructions was then performed on both the point and line source phantoms, with the aim of producing images with varying spatial resolution. This will assess each phantom's ability to demonstrate the effects of altering reconstruction parameters. Table 7.3 details the parameters used in each case.

Reconstruction Name	Reconstruction Parameters
Low Resolution	Plain HD, 18 effective iterations, 10mm Gaussian filter, Heavy z-axis filter
Medium Resolution	TOF+PSF, 54 effective iterations, 4mm Gaussian filter, Standard z-axis filter
High Resolution	PSF, 90 effective iterations, 0mm Gaussian filter, No z-axis filter

Table 7.3: Reconstruction parameters used to vary spatial resolution

7.3.4 Image Analysis

All line profile and ROI measurements were performed using Hermes Medical Systems' Hybrid Viewer software. Line profiles were fitted to Gaussian functions using ImageJ curve-fitting software, and the FWHM of the Gaussian fit was measured [53], [206]–[208]. It is generally accepted that a source should be sampled by at least three voxels in order to produce reliable Gaussian curve fitting: FWHM results obtained when less than three voxels sampled the sources are indicated where applicable, as these results may be unreliable.

7.3.4.1 Line Source Phantom

Figure 7.2 illustrates the perturbation subtraction process used to isolate line sources from the surrounding activity. Hermes Medical Systems' Hybrid Viewer software was used to subtract Scan 1's reconstructed images (surrounding activity only) from Scan 2's reconstructed images (line sources plus surrounding activity). Line profiles were analysed using both the Scan 2 images (lines in surrounding activity) and the subtraction images (line sources in near-zero surrounding activity). Profiles were positioned using CT data, and then copied across to the PET data sets as shown in Figure 7.3(a). Both horizontal and vertical profiles were drawn through both line sources using transaxial images. As the line sources ran parallel to the z direction, spatial resolution in the z direction was not assessed. Measurements were made on the central transaxial slice and four slices either side (9 neighbouring transaxial slices in total) to provide an estimate of the mean and error of the measurement method (this method was previously used by Alessio et al [49]).

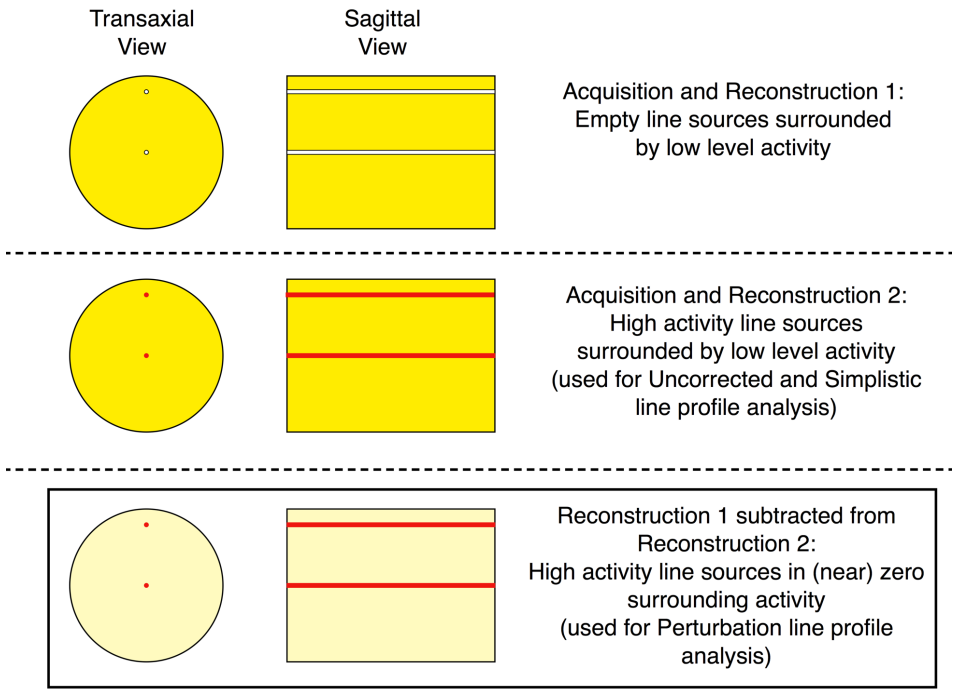


Figure 7.2: Perturbation method subtraction for line source's surrounding activity

Simplistic background correction was then applied to Scan 2's line profiles. A background value was obtained by placing a large ROI over an area of uniform surrounding activity, as shown in Figure 7.3(b). The ROI's mean voxel value was taken to be the 'background' activity concentration and subtracted from Scan 2's line profile data to produce simplistic background-corrected line profiles.

(a) **(b)**

Figure 7.3: Line source phantom (a) line profiles and (b) background ROI

Three horizontal (x-axis) and three vertical (y-axis) line profile data sets were therefore produced for both line sources within the phantom:

- **Uncorrected Method:** Line sources profiles in surrounding activity
- **Perturbation Correction Method:** Isolated line source profiles produced using perturbation subtraction
- **Simplistic Correction Method:** Line source profiles in surrounding activity corrected using simplistic background subtraction

7.3.4.2 Point Source Phantom

Point sources were analysed by drawing line profiles in the x, y and z directions, as shown in Figure 7.4. Line profile analysis was performed using the same methodology as the line source phantom, creating the following data sets:

- **Uncorrected Method:** Point source profiles in surrounding activity
- **Simplistic Correction Method**: Point source profiles in surrounding activity corrected using simplistic background subtraction.

(a) (b)

Figure 7.4: Point source phantom line profiles
Line profiles are positioned on (a) transaxial and (b) coronal PET images

7.3.4.3 Voxel Size Assessment

Horizontal line profiles were drawn through the ^{22}N point source (water phantom) and the central ^{18}F line source (200:1 phantom). Corrections for the surrounding activity in the ^{18}F line source phantom were performed using simplistic background subtraction only (this will be justified in the Results section on this chapter).

7.4 Results

7.4.1 Effects of Voxel Size on Spatial Resolution Measurement

The first aim of this chapter was to assess how spatial resolution measurement is affected by voxel size; in particular, to demonstrate at which point voxel size becomes a limiting factor upon FWHM measurement. Figure 7.5 plots FWHM measurements against voxel size for both the ^{22}Na point source (0.25mm diameter) and the ^{18}F line source (1.2mm diameter), when different effective iterations are applied.

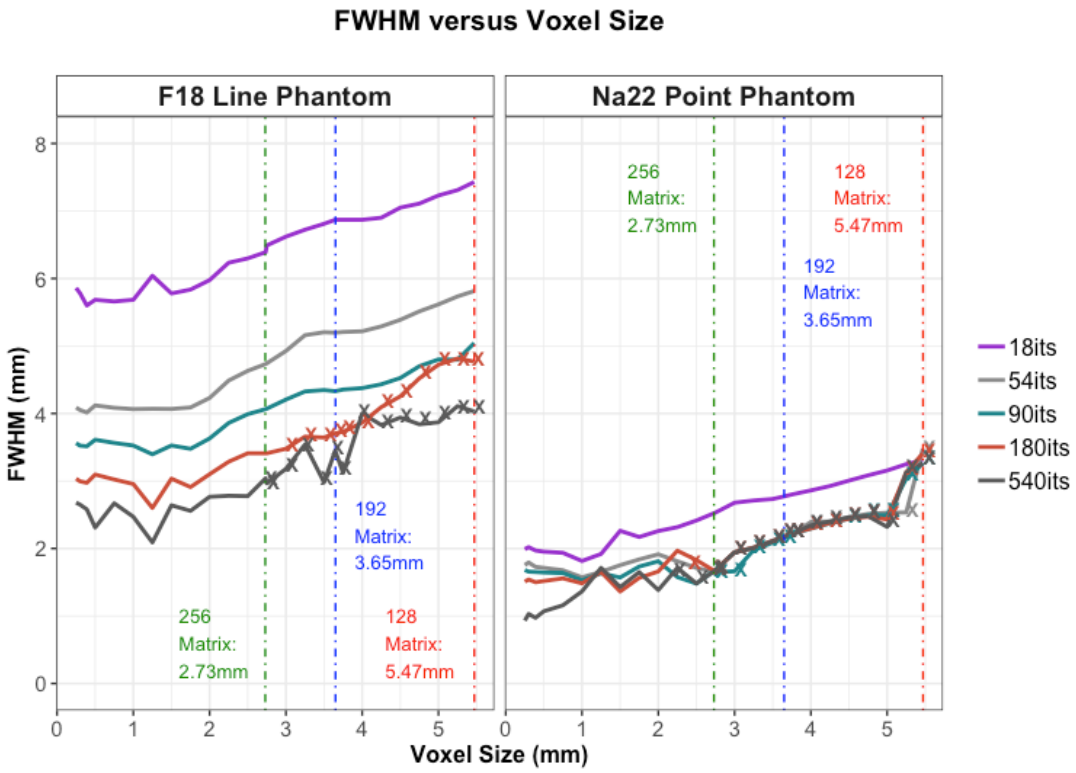


Figure 7.5: Spatial resolution versus voxel size as effective iterations vary

Both 200:1 ^{18}F line source and ^{22}Na point source in water results are shown. Crosses indicate where Gaussian curve fitting was unreliable due to inadequate source sampling. Dashed lines indicate voxel sizes achievable with full 700mm transaxial FOV.

The ^{22}Na point source FWHM results are much lower than those of the ^{18}F line source. One would expect this to be the case because of the point source's smaller size and much higher activity concentration. Both sets of results demonstrate that spatial resolution improves with increasing effective iterations, which is also as one would expect. Figure 7.5 also demonstrates that resolution measurements of both sources worsen as voxel size increases, which is also as expected.

The ^{22}Na point source resolution doesn't appear to degrade until the voxel size reaches approximately 3mm, when between 54 and 180 effective iterations are applied. However, the point source was sampled by less than three voxels (indicated by crosses in Figure 7.5) when the following conditions were met:

- Voxel size at least 4.75mm when 54 effective iterations are applied
- Voxel size at least 2.75mm when 90 effective iterations are applied
- Voxel size at least 2.50mm when 180 effective iterations are applied
- Voxel size at least 2.25mm when 540 effective iterations are applied

With respect to the 700mm FOV used for whole-body imaging, the 256 matrix achieved adequate sampling of the ^{22}Na point source when up to 90 effective iterations were applied.

The 192 matrix failed to achieve adequate sampling when more than 54 effective iterations were applied. The 128 matrix only achieved adequate sampling at the lowest iteration setting.

It should be noted, however, that three of the voxel sizes which provided inadequate sampling of the ^{22}Na source are smaller than the GEMS Discovery 690's z-direction voxel dimension, which is fixed at 3.34mm. This z-axis voxel sampling is used successfully in NEMA-style FWHM measurements, which use point sources with larger physical dimensions and smaller activity concentrations than the ^{22}Na source used in this experiment. This implies the ^{22}Na point source presents a greater challenge in terms of the voxel size required for adequate sampling. This is as expected: smaller objects require smaller voxels for adequate sampling (Nyquist Theorem).

The ^{18}F line source resolution doesn't appear to degrade until voxel size reaches approximately 2mm (for all effective iterations settings). However, the line source was sampled by less than three voxels when the following conditions were met:

- Voxel size at least 3.00mm when 180 effective iterations are applied
- Voxel size at least 2.75mm when 540 effective iterations are applied

The 256 matrix therefore achieved adequate sampling of the ^{18}F line source, even at high iterations. The 128 and 192 matrices failed to achieve adequate sampling once 180 effective iterations were applied.

One would expect the smaller ^{22}Na point source to achieve a smaller FWHM than the larger ^{18}F line source, and Figure 7.5 shows this is indeed the case for all voxel sizes. However, one would also expect that the voxel size required to degrade FWHM measurements would be smaller for the smaller source, due to Nyquist sampling theory. This is not the case for the results in Figure 7.5: the ^{18}F line source FWHM is degraded once voxels exceed approximately 2mm, whereas the ^{22}Na source FWHM is only degraded once voxels reach approximately 3mm. It is important to note the differences between the two phantoms extend beyond the physical size of the point/line source:

- The ^{22}Na point source activity was 0.525MBq, while the line source activity was 0.48MBq distributed along the full 20cm length of the line. The images of the line source therefore had greater statistical noise than the images of the point source.
- The ^{18}F line source was placed inside a background activity (200:1 line to background ratio), whereas the point source was placed within plain water. Reconstruction of the point source was therefore unaffected by the presence of any background activity.

It is therefore difficult to make a meaningful comparison between the two sets of results shown in Figure 7.5; the difference in critical voxels sizes may not be as pronounced in reality as it appears in these experimental results. However, the results indicate the use of the full 700mm FOV with 192 or 128 matrices (i.e. 3.65mm or 5.47mm voxels respectively) will compromise the reliability of Gaussian curve fitting, particularly for the smaller ^{22}Na source.

To ensure adequate sampling in a manner that is clinically relevant to liver imaging, the remainder of this chapter will use the full 700mm FOV with only the 256 matrix. This will provide reliable measurements for the ^{18}F line phantom at all effective iterations. FWHM measurements of the ^{22}Na source may be less reliable at higher iterations (180 and above) but are included for completeness.

7.4.2 Point Source Phantom

7.4.2.1 Effect of ^{18}F Correction Factors

The first aim of the ^{22}Na point source work was to determine if the use of ^{18}F correction factors affected FWHM measurements. Figure 7.6 shows medium resolution PET reconstructions of the ^{22}Na point source within a plain water background, using (a) ^{22}Na and (b) ^{18}F correction factors, with 1-voxel-wide horizontal line profiles. Images are zoomed to better visualise the point sources. Maximum voxel values are also included.

(a)
^{22}Na Correction Factors
Max Voxel = 3,874,468Bq/ml

(b)
^{18}F Correction Factors
Max Voxel = 3,618,722Bq/ml

Figure 7.6: Comparison of point source images acquired using (a) ^{22}Na and (b) ^{18}F correction settings

Images have been zoomed to better visualise point source (identical zoom factors applied to each acquisition). Images are scaled to maximum voxel values. Line profile width is 1 voxel.

A qualitative comparison of the ^{22}Na and ^{18}F corrected point source images demonstrates little difference between the two reconstructions. Maximum voxel values were within 6.6% of each other. Line profiles were then used to calculate the FWHMs: Figure 7.7 compares the raw line profiles (left) and the Gaussian fit (right) of both the ^{22}Na-corrected and ^{18}F-corrected line profiles ($R^2 > 0.99$ in both cases).

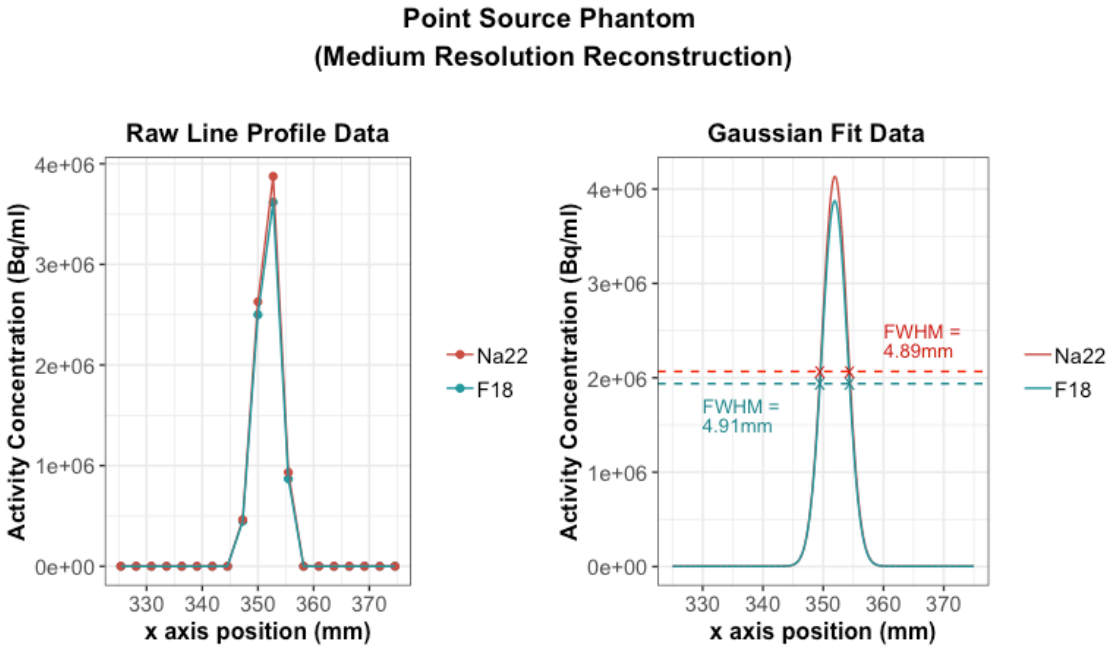


Figure 7.7: Raw line profile data and Gaussian fit data for point source phantom Data acquired using both ^{22}Na and ^{18}F correction settings.

Figure 7.7 demonstrates that both the raw data and the Gaussian fit data are similar for both isotope correction methods. Maximum values of the line profile data are consistent with maximum image voxels shown in Figure 7.6, demonstrating that both line profiles successfully sampled the maximum voxel. Furthermore, the measured FWHM of the Gaussian fit data are almost identical: 4.89mm for the ^{22}Na corrected data and 4.91mm for the ^{18}F corrected data (0.41% difference). This data suggests the use of ^{18}F correction factors for ^{22}Na point source reconstructions does not affect the point source FWHM. The assessment was then extended to include measurements of the x, y and z-axes for low, medium and high resolution reconstructions, as summarised in Table 7.4.

	x-axis			y-axis			z-axis		
	^{22}Na	^{18}F	Diff	^{22}Na	^{18}F	Diff	^{22}Na	^{18}F	Diff
Low-Res	10.17	10.18	0.01 (0.1%)	10.38	10.39	0.01 (0.1%)	7.51	7.49	0.02 (0.27%)
Med-Res	4.89	4.91	0.02 (0.41%)	4.26	4.26	0.0 (0.00%)	4.90	4.89	0.01 (0.20%)
High-Res	1.67	1.67	0.00 (0.00%)	1.79	1.79	0.00 (0.00%)	2.18	2.21	0.03 (1.38%)

Table 7.4: Point source spatial resolution measurements using ^{22}Na and ^{18}F correction factors

Low, medium and high resolution reconstructions were analysed. Magnitudes of the differences between ^{22}Na and ^{18}F corrections results are shown. Percentage differences with respect to ^{22}Na results are also shown.

Table 7.4 further demonstrates there is little difference in FWHM measurements when comparing ^{22}Na and ^{18}F correction factors: all 9 pairs of results are within 1.4% of each other, with 8 of the pairs within 0.5% of each other. Although a limited range of reconstructions was assessed, this experiment indicates the use of ^{18}F correction factors does not affect the measured FWHM of a ^{22}Na point source, and therefore fulfils the first aim of the ^{22}Na phantom work. Further phantom work was therefore undertaken by placing the ^{22}Na point source within a background volume of ^{18}F activity, and reconstructing the data using ^{18}F correction factors.

7.4.2.2 Effect of Increasing Background Activity

The second aim of the ^{22}Na point source phantom work was to determine if placing the source within clinically relevant activity concentrations of ^{18}F had any effect on FWHM measurements. The phantom background chamber was filled with five different ^{18}F activity concentrations, as described in Table 7.1. Phantom images are shown in Figure 7.8: the top row shows images scaled to a maximum SUV of 7 (threshold used routinely for viewing clinical images). The bottom row images are scaled to each image's maximum voxel value with a zoom factor applied to better visualise the point source.

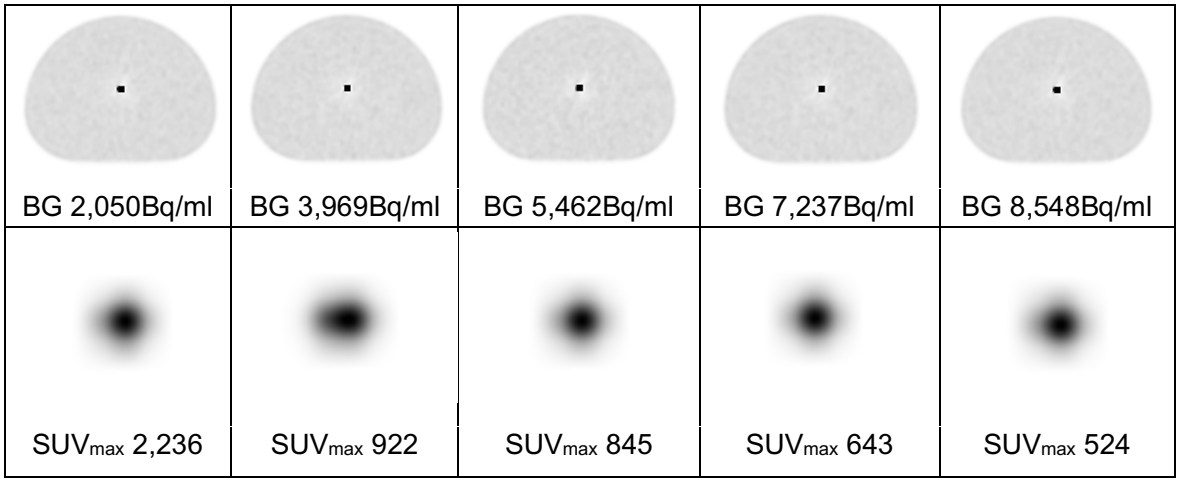


Figure 7.8: Point source phantom images with different source-to-background ratios

Medium resolution reconstruction parameters. Images on top row are scaled to maximum SUV_{max} = 7, as used for clinical image viewing. Images on bottom row are scaled to maximum voxel value and zoomed to better visualise point source.

Background activity is not visualised in the bottom row images because of the very high source-to-background ratios. It is difficult to visualise any differences in background activity in the top row images for the same reason. These images also appear to demonstrate some ringing artefacts in the immediate vicinity of the point source, although this may also be a result of the 'cold' Lucite surrounding the point source.

Figure 7.8 demonstrates qualitatively that there is little effect on the point source width as a result of increasing background radioactivity. The second point source measurement (SUV_{max} = 922) appears to have an elongated width along the x-axis, but as this effect is not observed on any of the other acquisitions, it is likely to be caused by a positioning error. Point source FWHMs in all three planes were then measured quantitatively using line profiles. Simplistic background correction was performed for each acquisition to correct for the different surrounding activity concentrations (the use of the simplistic background correction will be justified later in this chapter). Figure 7.9 plots the measured FWHM in all three directions against increasing background radioactivity concentration. Both uncorrected and simplistic corrected results are shown. Linear trendlines with R^2 values are also shown for each set of measurements.

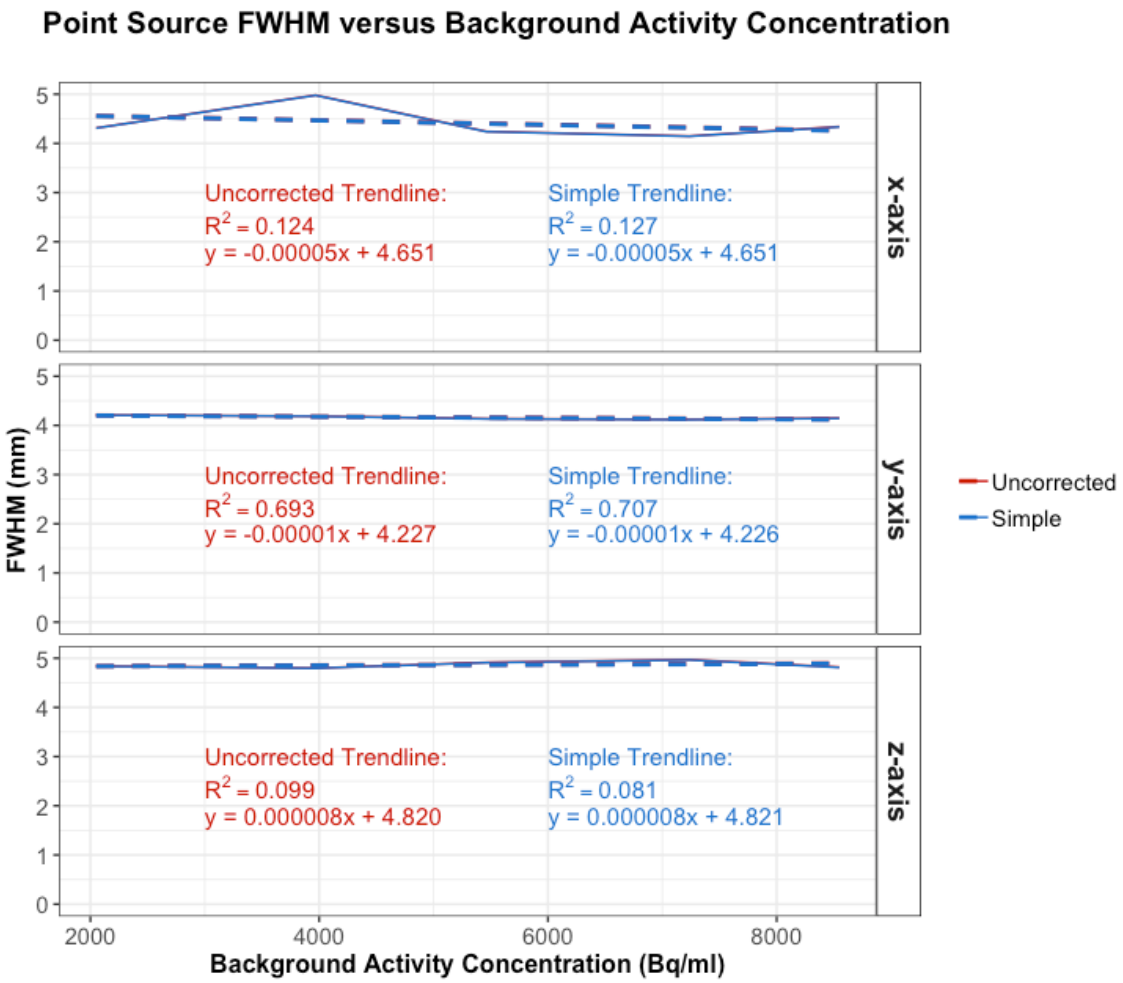


Figure 7.9: Spatial resolution versus background activity concentration

Results are shown for x, y and z directions. Both uncorrected (red) and simplistic corrected (blue) data are shown. Analysis was performed on medium resolution reconstruction images.

Figure 7.9 demonstrates that increasing background activity has little effect on FWHM. Linear trendlines demonstrate a poor fit and are not statistically significantly different to a

zero gradient (p-values from linear regression significance tests are above 0.05 in each case). These results therefore do not demonstrate any correlation between FWHM measurements and background activity. Furthermore, the Uncorrected and Simple results are almost identical; any discrepancies between the two correction methods are highly unlikely to be of clinically significance. One may therefore conclude that background correction is not required for the particular combination of point source and background activity concentrations used in this experiment. Background correction may be required at lower activity concentration ratios, but ratios in the order of 100:1 are unachievable with this particular point source activity (a 100:1 ratio would require approximately 6,400GBq in the background volume, which is both clinically and physically unrealistic).

This experiment therefore fulfils the second aim of the ^{22}Na phantom work. The activity concentration of the surrounding activity does not affect the FWHM measurements of the high activity point source when clinically relevant activity concentrations are used. Therefore, this phantom is not useful for assessing the effects of different lesion-to-background ratios due to the very high activity concentration of the point source. Furthermore, simplistic background correction did not affect the FWHM results, suggesting that background correction is not required for the range of background activities examined in this experiment: this is also due to the high activity of the point source. This partially fulfils the third aim of the ^{22}Na phantom experiment, which is to assess the requirement for background correction.

7.4.2.3 Effect of Altering Reconstruction Resolution

The fourth and final aim of the ^{22}Na phantom work was to assess the phantom's ability to demonstrate the effects of altering reconstruction parameters. Three different resolution reconstructions were performed for the ^{22}Na phantom with the highest ^{18}F background activity, as shown in Figure 7.10. Images are shown on two image scales as before, with images zoomed to better visualise the point source.

Qualitative image analysis demonstrates obvious differences in resolution, as one would expect. The point source appears large in size and is surrounded by a more homogenous background activity region on the low resolution reconstruction. The medium resolution reconstruction has a less homogenous background activity but has a sharper point source image. The high resolution phantom has the sharpest point source and the most heterogeneous surrounding activity of the three reconstructions. On the images scaled to a maximum SUV of 7, the high resolution phantom has a more pronounced 'cold' area surrounding the point source than the medium resolution phantom. Cold voxels on the medium and high resolution reconstructions are only visualised here because the maximum

scaling factor is relatively low compared to the maximum SUVs in the images. These cold voxels have no measurable effect upon FWHM analysis.

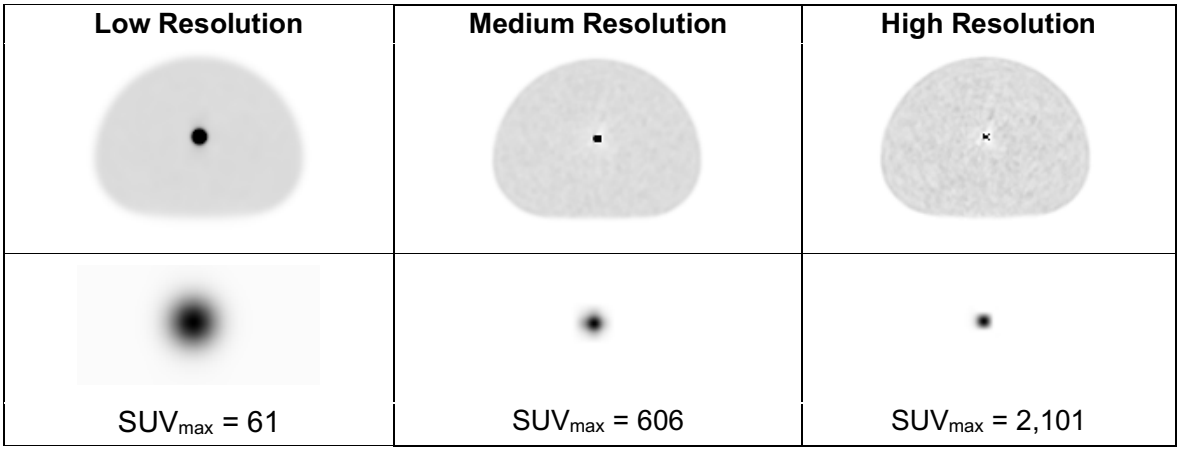


Figure 7.10: Low, medium and high resolution reconstructions of point source phantom within surrounding activity

Images on top row are scaled to maximum SUV_{max} = 7, as used for clinical image viewing. Images on bottom row are scaled to maximum voxel value and zoomed to better visualise point source. Surrounding activity concentration is 8,548Bq/ml.

Line profile analysis results (both Uncorrected and Simplistic Correction) for x, y and z axes are shown in Table 7.5.

	x-axis			y-axis			z-axis		
	Uncor	Simple	Diff	Uncor	Simple	Diff	Uncor	Simple	Diff
Low Res	10.733	10.578	0.155 (1.44%)	10.959	10.808	0.151 (1.38%)	7.545	7.442	0.103 (1.37%)
Med Res	4.342	4.334	0.008 (0.18%)	4.149	4.146	0.003 (0.07%)	4.819	4.811	0.008 (0.16%)
High Res	2.086	2.084	0.002 (0.10%)	2.262	2.261	0.001 (0.04%)	2.641	2.639	0.002 (0.08%)

Table 7.5: Spatial resolution measurements of point source surrounded by activity

Low, medium and high resolution reconstructions were analysed. Results are shown for x, y and z directions. Magnitudes of the differences between uncorrected ('Uncor') and simplistic ('Simple') corrections results are shown. Percentage differences with respect to uncorrected results are also shown.

As with the results in the previous section, differences between Uncorrected and Simple data are small. The magnitudes of these differences are dependent upon image resolution: maximum differences were 0.1% for high resolution, 0.18% for medium resolution and 1.44% for low resolution. These small differences are unlikely to be of clinical significance. This experiment therefore fulfils the third aim of the ^{22}Na phantom work: simplistic background correction does not have a clinically significant effect on FWHM results,

suggesting that background correction is not required for the combination of ^{22}Na source activity and background activities examined in this experiment.

Table 7.5 also demonstrates clear differences in FWHMs achieved by the three different resolutions, as one would expect from the qualitative assessment. Medium resolution FWHMs are approximately double those of the high resolution, while the low resolution FWHMs are approximately double those of the medium resolution. This therefore fulfils the fourth and final aim of the ^{22}Na phantom experiment. The ^{22}Na phantom can be used to demonstrate the relative effects of altering reconstruction parameters, both qualitatively and quantitatively.

7.4.3 Line Source Phantom

The line source phantom was used to assess 'uncorrected', 'perturbation' and 'simplistic' corrections for surrounding activity, as well as the phantom's ability to demonstrate the effects of altering reconstruction parameters. The effects of two different source-to-background ratios were also assessed.

7.4.3.1 Qualitative Comparison of Corrected and Uncorrected Phantoms

Figure 7.11 shows transverse and sagittal images of both the 200:1 and 100:1 phantoms reconstructed with low, medium and high resolution. Images acquired before and after the line sources were filled are shown. Perturbation subtraction images are also shown. Qualitative image analysis demonstrates the relative spatial resolution performance of each reconstruction is as expected for both the 200:1 and 100:1 phantoms: as resolution increases, line sources appear sharper and surrounding activity appears less homogenous. Line sources are still visible following subtraction of surrounding activity for all three reconstructions; however, they appear to have reduced intensity following the subtraction. This was verified by performing a brief quantitative analysis of the line sources' maximum activity concentrations on the central transaxial slice before and after perturbation subtraction, as shown in Table 7.6. It should also be noted that the empty line sources are not visualised on the images showing surrounding activity only: there are no visible 'cold' areas coinciding with the positions of the empty line sources.

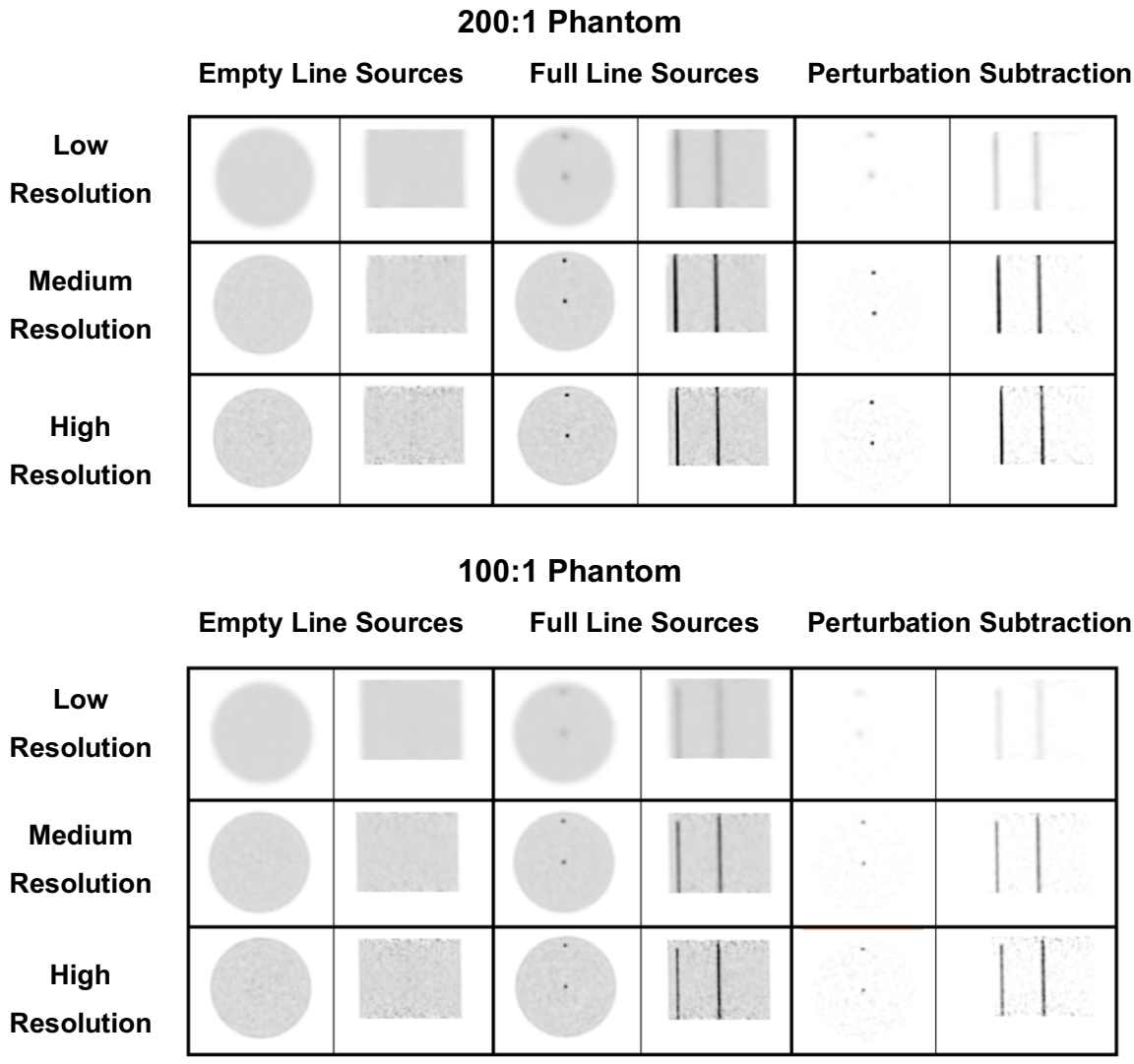


Figure 7.11: Low, medium and high resolution reconstructions of line source phantoms

Both 200:1 (top) and 100:1 (bottom) ratio phantoms shown. Surrounding activity with empty line sources (left), uncorrected line sources within surrounding activity (centre) and line sources following perturbation method subtraction of surrounding activity (right). Images scaled to maximum SUV = 7.0, as used for clinical image viewing.

Table 7.6 confirms the perturbation method produces line sources of reduced intensity. Lower resolution images demonstrated the greatest reduction in maximum activity concentrations, while the 100:1 phantom demonstrated greater reductions in maximum activity concentrations than the 200:1 phantom. Furthermore, the peripheral line source demonstrated a smaller percentage reduction than the central line source as a result of the perturbation subtraction.

		Maximum Activity Concentration (Bq/ml)					
		Low Resolution		Medium Resolution		High Resolution	
		Before	After	Before	After	Before	After
200:1 Ratio	Periphery	16,069	8,005 (-50%)	53,027	46,352 (-13%)	76,020	69,702 (-8%)
	Central	16,159	7,799 (-52%)	53,125	45,373 (-15%)	74,270	64,847 (-13%)
100:1 Ratio	Periphery	11,952	4,004 (-66%)	29,760	22,445 (-25%)	39,483	31,572 (-20%)
	Central	12,161	3,819 (-69%)	30,249	21,027 (-30%)	42,011	29,486 (-30%)

Table 7.6: Maximum activity concentration measurements of line sources

Results shown for both 100:1 and 200:1 phantoms, and for both uncorrected images ('Before') and perturbation images ('After'). Measurements made using central transaxial slice. Percentage reductions as a result of subtraction are shown in parentheses.

7.4.3.2 Quantitative Comparisons of Contrast Ratios and Corrections

Quantitative line profile analysis was then performed on the Uncorrected, Perturbation and Simplistic correction data sets, for all three resolution reconstructions and for both 200:1 and 100:1 phantoms. Results for the central line source using horizontal profiles are shown in Figure 7.12 (all plots use the same x and y axes scales to aid comparison between different contrast ratios and resolutions). No FWHM measurement could be obtained for the uncorrected low-resolution data, as the surrounding activity voxel values exceed half of the line source maximum voxel. Perturbation and simplistic correction methods produce similar line profiles and FWHM measurements. This is in contrast to what one may have expected, as illustrated by Figure 7.1: the simplistic correction was expected to produce narrower FWHM results than the perturbation method. However, Figure 7.1 assumes the data has greater spatial resolution than observed in the reconstructed images. As the empty line sources are not visualised on the images, even on the high resolution reconstruction, the perturbation correction method is in fact similar to the simplistic correction method at clinically available levels of image resolution.

Figure 7.12 also demonstrates that as resolution increases, peak values increase and the FWHMs decrease for all three correction methods, as one would expect. Peak values for the 200:1 phantom are greater than those of the 100:1 phantom, which is also as expected. The plots also demonstrate that as resolution improves, line profiles and FWHM results of all three correction methods become more similar to each other. For all three resolutions, and for both contrast ratios, the 'Perturb' and 'Simple' results are similar in terms of both

peak values and FWHMs, although this limited data suggests that the perturbation method produces marginally narrower FWHM results.

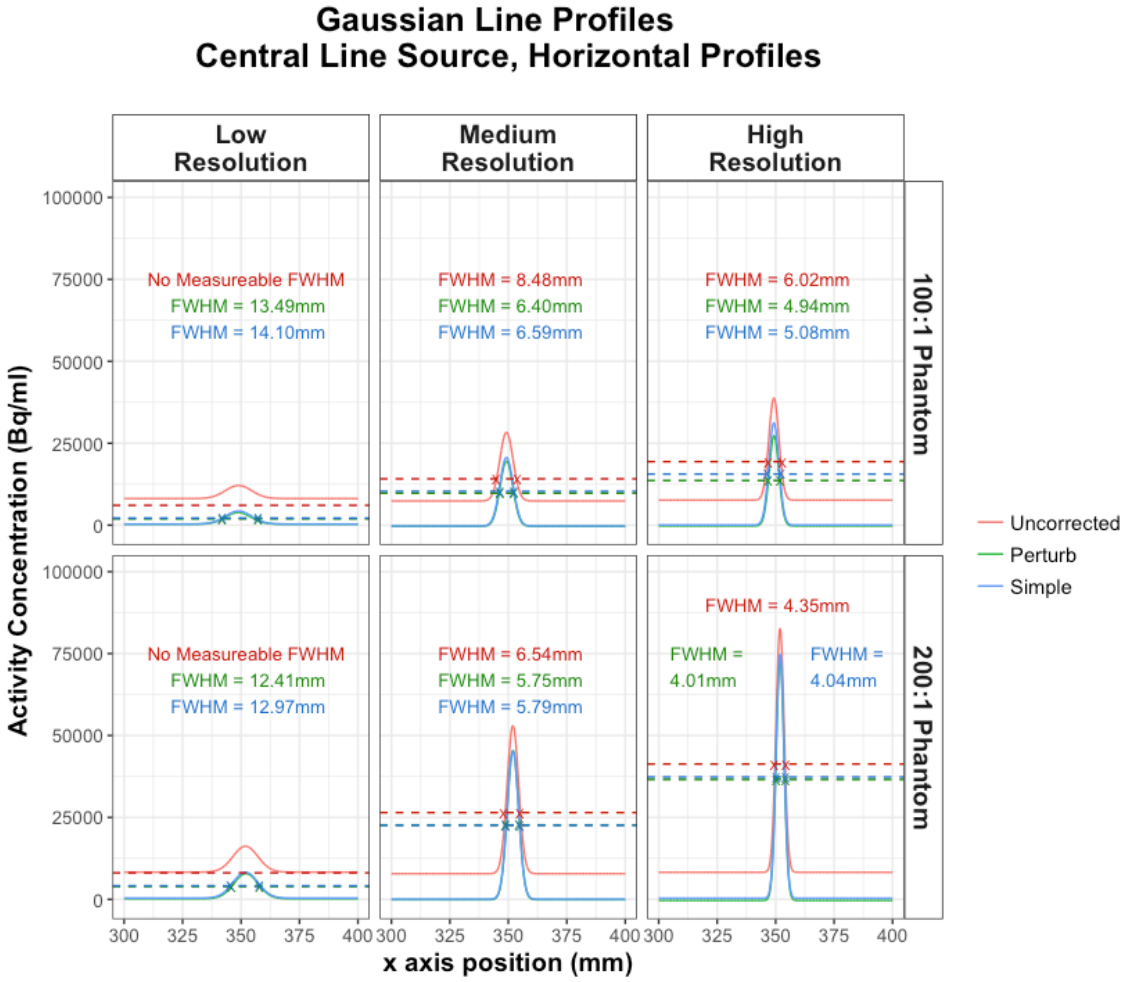


Figure 7.12: Gaussian fit data for line source phantoms (x-axis only)

Results are shown for uncorrected, perturbation corrections and simplistic corrections, and for both 100:1 and 200:1 phantoms. Results for all three resolutions are shown. Data is for horizontal profile through central line source on the central transaxial slice. Dashed lines denote half maximum values; crosses denote FWHM measurement points (where possible). FWHM results for Gaussian fit data are also shown (where possible).

FWHM analysis of the centre and peripheral line sources, using nine consecutive transaxial slices for each phantom reconstruction was then performed. Figure 7.13 shows the mean FWHM results and their 95% confidence intervals.

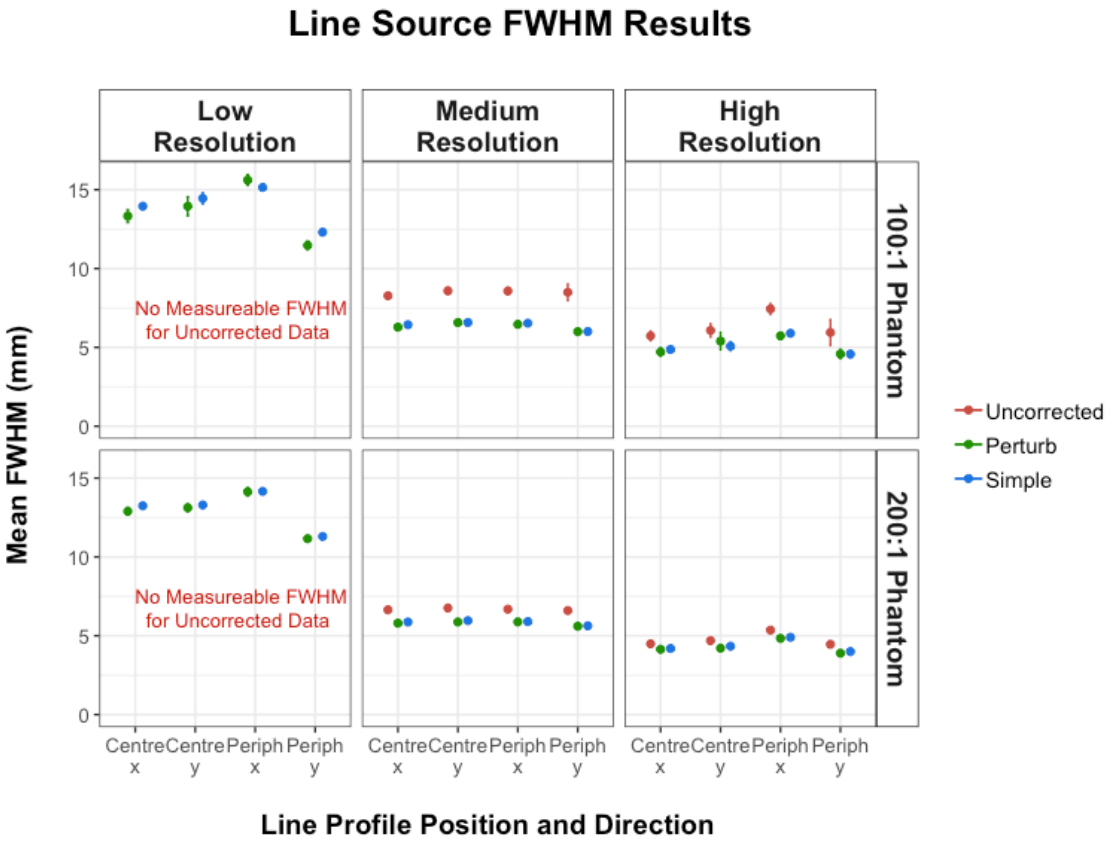


Figure 7.13: Mean spatial resolution results for line source phantoms

Results for uncorrected, perturbation correction and simplistic correction, for 100:1 and 200:1 phantoms, for all three reconstruction resolutions. Mean resolution results for both central and peripheral line sources are shown for both x and y axes. Error bars represent 95% confidence intervals.

Figure 7.13 confirms the observations made from Figure 7.12 apply to both the centre and peripheral line source, and both x and y axis directions:

- The 200:1 ratio phantom produces smaller FWHMs than the 100:1 phantom when the same reconstruction and correction methods are applied
- No FWHMS could be measured for uncorrected low resolution data: this demonstrates that the 'Uncorrected' method is not a valid approach to FWHM measurement.
- 'Perturb' and 'Simple' results appear to be similar for each measurement. Paired t-tests were performed on each set of results. 17 out of the 24 sets of measurements had p-values below 0.05, indicating that the differences between the two correction methods are statistically significant; however, as the differences are small in magnitude and do not alter the relative effects of the reconstruction resolution, these differences are highly unlikely to be of clinical significance.
- The centrally positioned line source has similar FWHMs in both the x and y axis directions, for all three resolutions

- The peripheral line source has greater FWHMs along the x-axis than the y-axis for all three correction methods and for both contrast ratios. This effect will be analysed in more detail in the next chapter.

The first aim of the line source phantom experiment was to compare the three methods used to correct for surrounding activity and determine which method should be used when performing comprehensive spatial resolution analyses. Figures 7.12 and 7.13 demonstrate the 'Uncorrected' method is not suitable for assessing low resolution reconstructions and can therefore be excluded from any further work. This therefore fulfils the first aim of the line source experiment: corrections for surrounding activity are required for spatial resolution measurement as 'Uncorrected' results cannot provide FWHM measurements for low resolution reconstructions. Furthermore, 'Perturb' and 'Simple' correction techniques produced similar FWHM results.

The second aim of the line source phantom experiment was to determine if the line phantom was able to demonstrate the effects of altering various reconstruction parameters. Figures 7.12 and 7.13 demonstrate that both 'Perturb' and 'Simple' methods differentiate between different levels of image resolution. The second aim of the line source experiment is therefore fulfilled: this phantom methodology can be used to assess the effects of altering reconstruction parameters.

The third aim of the line source phantom experiment was to determine if the use of a single contrast ratio could be sufficient to assess the effects of reconstruction parameters on spatial resolution. The 200:1 profiles have higher peak values than the 100:1 profiles, as one would expect from the higher activity concentration. The 200:1 profiles therefore produced FWHMs that were consistently smaller than those of the 100:1 phantom (average difference was 0.71mm). However, both contrast ratios clearly demonstrated the effects of reconstruction parameters, as shown in Figures 7.12 and 7.13. The third aim of the line source experiment is therefore fulfilled: whilst the contrast ratio has been shown to have an effect on the measured spatial resolution, a single ratio should be sufficient to demonstrate the effects of reconstruction parameters upon spatial resolution.

7.5 Discussion

The first experiment performed in this chapter assessed the effects of increasing voxel size upon FWHM measurements using both an ^{22}Na source in water and ^{18}F line sources within background activity. Results demonstrated that, for the particular imaging conditions used in this experiment, voxel sizes of up to 2mm for the ^{18}F line source and 3mm for the ^{22}Na

source could be used without degrading the FWHM measurements. In terms of devising a method to measure clinical spatial resolution when the full 700mm FOV is used, the results demonstrated that 128 and 192 matrices did not provide adequate sampling of either source, particularly at higher iterations. The remainder of spatial resolution measurements performed in this chapter therefore used the 256 matrix with the 700mm FOV; this provided reliable FWHM measurement of the line source phantom at all effective iterations assessed in this chapter. However, the results also indicated that even the 256 matrix did not provide adequate sampling of the ^{22}Na source when 180 effective iterations or more were applied.

The use of the ^{22}Na point source and ^{18}F line source phantoms for measuring spatial resolution in a clinically relevant manner were then assessed. Both phantoms had advantages and disadvantages.

The ^{22}Na point source phantom provides a long-lived reproducible point source that allows FWHM measurements to be made in all three planes. As demonstrated by Section 7.4.2.3, the phantom can be used to assess the effects of altering reconstruction parameters upon spatial resolution. However, the point source is a 0.25mm diameter, 0.525MBq ^{22}Na spherical point source. The activity concentration within the sphere was therefore approximately 64GBq/ml. Considering the limit for a whole-body 18F-FDG scan in the UK (as defined by the Administration of Radioactive Substances Advisory Committee (ARSAC)) is 400MBq, the activity concentration of this point source is likely to be many orders of magnitude greater than one would expect to observe within a patient. Furthermore, a consequence of the high activity concentration of the point source was that a clinically relevant source-to-background ratio could not be achieved realistically (a 100:1 ratio would require approximately 6,400GBq in the background volume). When the surrounding activity concentration was similar to that found within the liver of patients injected with 400MBq ^{18}F-FDG, the theoretical source-to-background ratio was approximately 7,500,000:1. The high activity concentration of the point source therefore eliminated the effects of the surrounding activity in the background cavity. The high activity of the point source is also likely to have caused Gibbs artefacts, although these are unlikely to have affected the FWHM measurements. Additionally, as only one point source was available, spatial resolution could only be measured in a single position for each phantom acquisition.

However, it should be noted that images of the 7,500,000:1 source-to-background ratio, reconstructed using typical clinical parameters, produced a voxel ratio in the order of 500:1. The ^{22}Na phantom method using ^{18}F surrounding activity may therefore be worth revisiting when the long-lived point source has decayed to a lower activity. Furthermore, the point source phantom could be improved by obtaining multiple point sources so resolution can

be measured at different positions within the FOV in a single acquisition. NEMA spheres could be filled and added to the phantom to assess how resolution is affected by the presence of additional objects within the FOV with activity concentrations greater than background.

There are several advantages of the line source phantom. The line sources and background cavity can be filled to achieve a range of clinically relevant contrast ratios. The phantom includes two line sources, which allows resolution to be measured in two physical positions for each phantom acquisition. Furthermore, the presence of two sources within the phantom also provides a slightly more complex imaging scenario, which is closer to the clinical situation, than a phantom containing a single source within background activity.

However, the line source phantom has disadvantages. Line sources cannot be used to assess spatial resolution in all three planes simultaneously: only the x and y directions were assessed in this chapter. Furthermore, fillable line sources are not as reproducible as the point source phantom, due to potential measurement, filling and positioning errors. The finite size of the line source (1.2mm diameter) was relatively large compared to the point source used in this chapter (0.25mm). Smaller diameter sources provide more accurate measurement of the imaging system's point source response function.

Three different methods of correcting for the surrounding activity were also assessed in this chapter. No corrections were required for the point source phantom, as the high contrast ratio rendered the surrounding activity effectively absent. However, corrections were required for the line source phantoms as these used lower, clinically relevant contrast ratios: this was demonstrated by the low-resolution phantom having no measurable FWHM when no corrections were applied.

The line source phantom results were used to compare the perturbation method with the simplistic correction method (the perturbation method was not used with the point source phantom, as explained in Section 7.3.1.1). Both correction methods provided a set of results that were not only similar, but clearly demonstrated the differences between the different reconstruction resolutions and the differences in contrast ratios.

The overall aim of this chapter was to recommend a methodology for measuring the relative effects of altering reconstruction parameters upon spatial resolution in a clinically relevant manner. Analysis has demonstrated the line source phantom to be the most appropriate phantom as it can be filled with clinically relevant contrast ratios and can measure resolution in two locations simultaneously. Although the line source can be filled with different contrast ratios, Figure 7.13 demonstrated that the use of a single contrast ratio is sufficient for

assessing the relative effects of altering reconstruction parameters. The perturbation and simplistic corrections produced similar results which are unlikely to demonstrate any clinically significant differences. As the simplistic correction method doesn't require additional acquisitions with empty line sources, it is less labour intensive than the perturbation method: the simplistic correction method is therefore more compatible with a large-scale analysis of spatial resolution.

7.6 Conclusions

The following phantom method is proposed for measuring the relative effects of altering reconstruction parameters in a clinically relevant manner:

- Line phantom with 200:1 ratio
    - This provides a clinically relevant contrast ratio, with easily visualised line source images (even at low resolution reconstructions).
- Simplistic corrections applied to account for surrounding activity
    - This technique is the fastest to implement as background-only acquisitions, reconstructions and subtractions are not required. A wider range of reconstructions can therefore be examined in the limited time available for this study.
- 700mm FOV, 256 matrix, 2.73mm voxel sizes in x and y directions
    - Using the maximum FOV means results will be applicable to liver imaging of larger patients. This also ensures line sources will be sufficiently sampled to allow reliable Gaussian curve fitting and FWHM measurement.

The next chapter will therefore use this methodology to assess how spatial resolution is affected when the various reconstruction parameters are altered.

Chapter 8 : Assessment of Spatial Resolution in Reconstructed Images of a Line Source Phantom

Whilst ^{18}F-FDG PET has high sensitivity and high accuracy for diagnosing hepatic and extrahepatic metastases, it has limited spatial resolution when compared to high-resolution modalities such as CT and MRI [209]. Limited spatial resolution, and the associated partial volume effects, is particularly disadvantageous when attempting to identify small lesions within the liver [210], as it causes blurring in reconstructed images (as discussed in Chapter 1). Small lesions appear larger in size and lower in intensity, causing SUVs to be underestimated [76]. Small, low contrast lesions within the liver are also challenging to identify due to the mottled, moderate uptake within the surrounding liver tissue.

The primary aim of this thesis is to develop a generic methodology to assess and optimise PET image reconstruction whilst considering clinical context, whilst the secondary aim is to use this methodology to optimise reconstruction on the GEMS Discovery 690 for the detection of small liver lesions. In order to optimise PET reconstruction for this purpose, it is important to characterise the effects of the various reconstruction parameters upon spatial resolution. When combined with the assessments of noise (Chapter 6) and lesion detection (Chapter 9), this will enable the selection of a combination of reconstruction parameters to produce the optimal trade-off between spatial resolution and other aspects of image quality, in order to maximise the accuracy of liver lesion detection.

The previous chapter demonstrated that clinical spatial resolution is challenging to assess, and as a result, there is no established method for its assessment in the literature. The chapter therefore evaluated different phantom and analysis techniques to assess spatial resolution in a clinically relevant manner. The chapter concluded by recommending the use

of a line source phantom and applying a simplistic correction method to account for the background activity. This methodology was shown to demonstrate how different reconstruction parameters affected clinical spatial resolution when a single source-to-background ratio was used. This chapter uses this methodology to perform a more comprehensive assessment of how clinical spatial resolution is affected by different combinations of reconstruction parameters.

8.1 Introduction

The effects of both reconstruction iterations and post reconstruction filtering upon spatial resolution have been well reported in isolation: increasing effective iterations improves spatial resolution (at the expense of increased noise) and increased filtering degrades spatial resolution (while reducing image noise). However, the effects on clinical spatial resolution as a result of PSF and TOF (in combination with effective iterations and filtering) are less certain, despite several studies undertaking investigations. This is partly due to the difficulties, and inter-study inconsistencies, of measuring spatial resolution in a clinically relevant manner, as discussed in the previous chapter.

The GEMS White Paper on 'SharpIR' [89], the GEMS implementation of PSF modelling, concluded that PSF modelling improves spatial resolution for a line source in air. Their methodology involved imaging an ^{18}F capillary tube in air at varying distances from the FOV centre. Four reconstructions were assessed: HD, PSF, TOF and PSF+TOF. Two PET-CT systems were used in this study: a GEMS Discovery 600 was used to compare HD and PSF, while a GEMS Discovery 690 was used to compare TOF and PSF+TOF. PSF reconstructions used 2.5 times more effective iterations than the non-PSF reconstructions, as PSF modelling is known to require more iterations to reach convergence. PSF effects were demonstrated by comparing HD against PSF results (Discovery 600), and TOF against PSF+TOF (Discovery 690). PSF and PSF+TOF FWHM measurements were approximately 2mm throughout the FOV, while HD and TOF worsened with increasing distance from the FOV centre. PSF improved spatial resolution by approximately 2mm (50% improvement) at the FOV centre, and by 4mm (67% improvement) at a distance of 250mm.

The GEMS White Paper therefore appears to suggest PSF modelling significantly improves spatial resolution performance, as measured by the FWHM of a line source throughout the FOV, when compared with HD and TOF reconstructions. However, there were major limitations to this study. The most significant limitation was the absence of scatter material and background activity (the importance of which was discussed in Chapter 7); results cannot be considered clinically relevant. Furthermore, PSF reconstructions used smaller

voxel sizes (0.65mm) than non-PSF reconstructions (0.98mm), which would likely result in lower FWHM results (particularly for sources in air [49]), and only the non-PSF reconstructions were subject to filtering (Gaussian 2mm FWHM filter).

The White Paper references a 2010 study by Alessio et al [49] as the publication that developed the 'SharpIR' PSF algorithm. The Alessio study assessed the effects of the PSF algorithm using multiple line sources both within air and within background radioactivity throughout the FOV. It should be noted that this study used an older PET-CT system (GEMS Discovery STE) that was not capable of TOF. The 'line source within air' results were similar to those reported in the GEMS White Paper: PSF produced significantly lower FWHM values than non-PSF reconstructions and were more consistent throughout the FOV. The 'line source within background radioactivity' results, however, demonstrated PSF FWHM values were in fact dependent upon the position within the FOV, although to a lesser extent than the non-PSF reconstruction. PSF demonstrated minor improvements over non-PSF at the FOV centre (approximately 3% FWHM reduction) but demonstrated greater improvements for the peripheral line source (approximately 12.5% FWHM reduction). The benefit of PSF was improved when the number of effective iterations was increased; however, this study noted that the PSF algorithm had not yet converged at FOV edges even after 2,800 effective iterations were applied. Almost 200 effective iterations were required to observe any improvement in FWHM as a result of PSF modelling for a line source positioned 5.1cm from the FOV centre, which are more effective iterations than typically used clinically. However, this study concluded that PSF improves spatial resolution when typical clinical levels of effective iterations are used (between 50 and 160 effective iterations), when the effects of image noise are also taken into account: PSF improved spatial resolution by approximately 15% at matched noise levels.

Despite Alessio's findings on the use of background radioactivity when assessing PSF, several more recent studies have drawn conclusions on PSF performance using point/line sources in air. A 2016 study by Murata et al [197] used an ^{22}Na source positioned at multiple points within the FOV to quantify the effects of PSF on two different PET-CT systems. No scatter material or background radioactivity was used, and no corrections for scatter or attenuation were applied during reconstruction. PSF on both scanners was relatively uniform throughout the FOV and demonstrated improvements over non-PSF reconstructions, particularly at the FOV edges. The study acknowledged that the lack of background radioactivity may have caused the PSF algorithm to over-correct the point source resolution. A 2015 study by Suljic et al [201] used three ^{18}F line sources in air to assess the impact of TOF and PSF on a Siemens Biograph PET-CT scanner. PSF was found to improve spatial resolution within 10cm of the FOV centre by up to 43%, whereas TOF was found to have a negligible effect on spatial resolution. They also concluded PSF

achieves uniform spatial resolution throughout the FOV. They state their results are in line with the results of other studies using similar methodology, and the limitations caused by the omission of background radioactivity are not discussed.

8.2 Aims

The literature review demonstrated few studies have assessed the effects of reconstruction parameters upon spatial resolution using clinically relevant methodology. In particular, the use of sources in air as opposed to background activity may have led studies to exaggerate the clinical benefits of PSF modelling.

The aims of this chapter are to characterise the effects of the following reconstruction parameters upon clinical spatial resolution, using the line source phantom methodology developed in the previous chapter:

a. Gaussian filter width
b. Effective iterations
c. Reconstruction method (TOF and PSF)
d. Consistency of resolution at different positions within the FOV

The effects of the z-axis filter are not explicitly assessed in this chapter, as the line source phantom is only used to assess resolution in the x and y axes directions.

The effects of varying voxel sizes on spatial resolution were assessed in the previous chapter. As a result, this chapter exclusively uses a 700mm FOV with a 256 matrix. This preserves the maximum FOV so results are applicable to larger patients, whilst ensuring adequate sampling of the line sources for reliable FWHM measurements.

Spatial resolution will be assessed both qualitatively and quantitatively. However, this chapter does not aim to assign a numerical value to the spatial resolution of the system, in a similar manner to NEMA spatial resolution tests. Instead, the qualitative and quantitative assessments will be used to assess the relative performance of different combinations of the reconstruction parameters. When combined with the assessments of noise (Chapter 6) and lesion detection (Chapter 9), this will enable the selection of a combination of reconstruction parameters to produce the optimal trade-off between spatial resolution and other aspects of image quality, in order to maximise the accuracy of liver lesion detection.

8.3 Materials and Methods

This chapter uses the methodology for assessing spatial resolution recommended by the previous chapter:

- Line phantom with 200:1 line-to-background ratio
- Simplistic correction applied to account for surrounding activity

8.3.1 Line Source Phantom and Acquisition Protocol

The NEMA 1994 line source phantom, containing two fillable line sources (diameter 1.2mm) within a 20cm diameter fillable cavity, was used in this chapter. This phantom is described in more detail in Section 3.2.3.2, and the acquisition used in this chapter is described in Section 7.3.2.2. The 200:1 source-to-background ratio, detailed in Table 8.1, was shown to clearly demonstrate the effects of varying reconstruction parameters on spatial resolution.

Surrounding Activity Concentration	Line Source Activity Concentration	Ratio	Line Image Maximum Activity Concentration	SUVmax
8,316Bq/ml	1,609,575Bq/ml	194:1	53,124Bq/ml	7.23

Table 8.1: Line phantom activity concentrations

Resultant SUVmax values when GEMS' suggested reconstruction parameters are applied are also included.

The phantom has one line source positioned centrally and one line source positioned 7.5cm radially (near the phantom's periphery).

8.3.2 Image Reconstruction Protocols

Table 8.2 summarises the reconstruction parameters used in this chapter. All possible combinations of the reconstruction parameters shown in Table 8.2 were used: 120 reconstructions were therefore performed in total.

Reconstruction Parameter	Settings Used	Number of Result Groups
Effective OSEM Iterations	18, 54, 90, 180, 540	5
Reconstruction Method	HD, PSF, TOF, PSF+TOF	4
Gaussian Filter (FWHM)	0 →10mm, in 2mm increments	6
Z-Axis Filter	None	1
Matrix Size	256*256	1

Table 8.2: Reconstruction parameters used for line source phantom

All reconstructions in this chapter were performed using the full 700mm transaxial FOV and a 256 matrix (voxel dimensions 2.73mm x 2.73mm x 3.34mm). This ensures results are applicable to larger patients, whilst providing adequate sampling of the line sources for reliable FWHM measurements (as demonstrated in the previous chapter).

As the phantom does not facilitate measurement of z-axis resolution simultaneously with x and y axes resolution, and the analysis required for each reconstruction was relatively time-intensive, only a single z-axis filter setting (no filter) was assessed in this chapter.

Although 540 effective iterations are unlikely to be used clinically (GEMS suggest using only 54 effective iterations), several studies have shown PSF requires more than 180 effective iterations to reach full convergence. Furthermore, Alessio et al [49] demonstrated approximately 200 effective iterations were required for PSF to improve resolution 5cm from the FOV centre; the phantom used in this chapter measures resolution at a distance of 7.5cm from the FOV centre. It is therefore prudent to include 540 effective iterations in this chapter in order to fully evaluate the effects of PSF upon spatial resolution.

8.3.3 Image Analysis

All line profile and ROI measurements were performed using Hermes Medical Systems' Hybrid Viewer software. Horizontal and vertical profiles were drawn through both the central and peripheral line sources on the central 9 transaxial slices. Background ROIs were also drawn on each of these slices and used to perform simplistic background correction (as described in the previous chapter). Corrected line profiles were then fitted to a Gaussian function using ImageJ curve-fitting software, and the FWHM of the Gaussian fit was calculated [53], [206]–[208].

(a) (b)

Figure 8.1: Line source phantom (a) line profiles and (b) background ROI

The nine FWHM results for each reconstruction were then used to calculate the mean result and 95% confidence intervals.

8.4 Results

8.4.1 Qualitative Analysis of Phantom Images

Figure 8.2 compares the effects of Gaussian filter width and effective iterations. All images shown were produced using the HD reconstruction method. Each row shows the results for a different Gaussian filter width, while each column shows the results for a different number of effective iterations (for brevity, only three of the iteration settings used are shown).

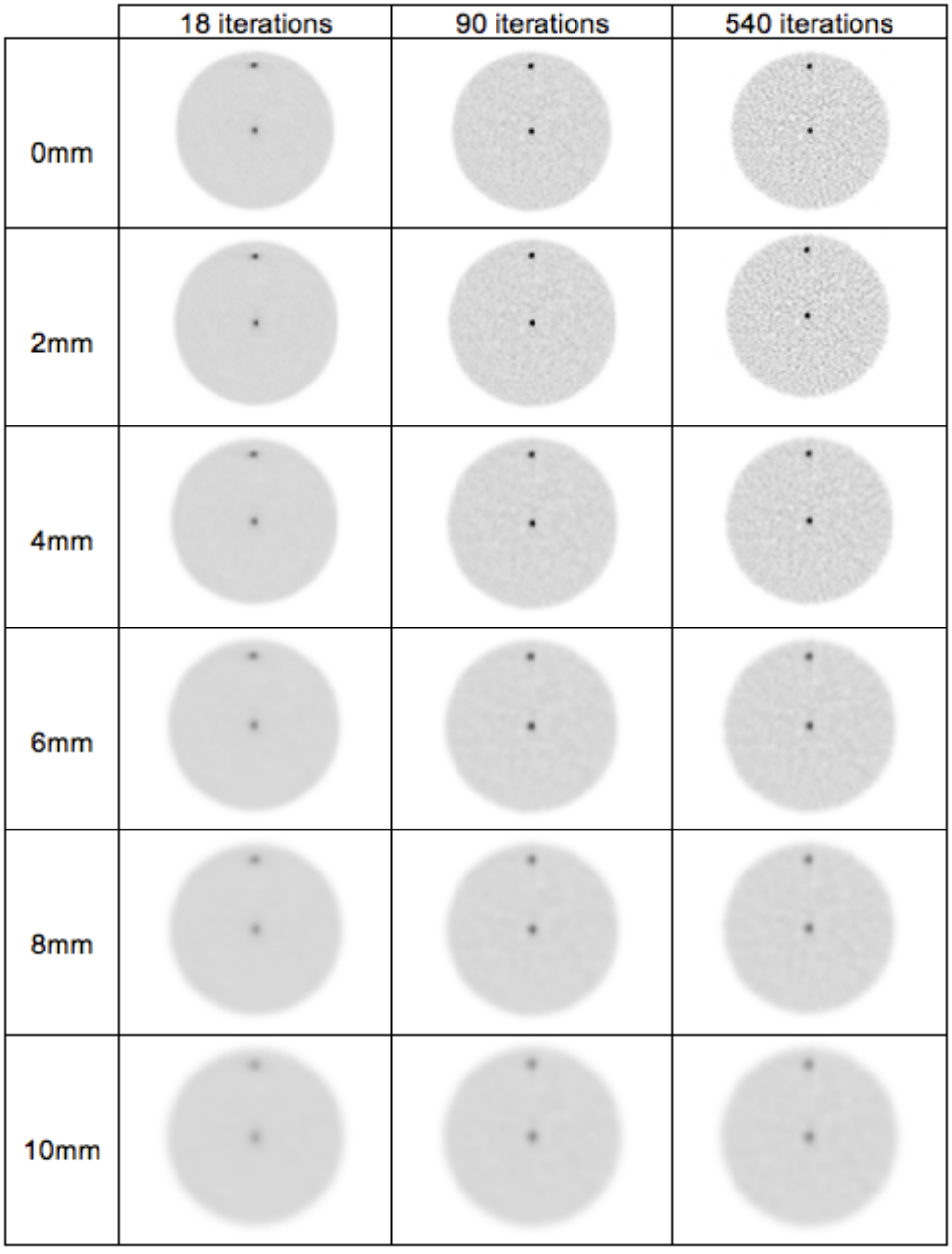


Figure 8.2: HD reconstruction images with varying filter width and effective iterations

Figure 8.2 generally demonstrates the expected trade-off between spatial resolution and image noise. Increasing filter widths degrade spatial resolution by introducing blurring, while

increasing effective iterations improves spatial resolution by producing noisier, sharper images. Figure 8.2 further demonstrates differences in resolution between the central and peripheral line sources. The peripheral line source in the 0mm, 18 iterations image appears to be blurred to a greater extent than the central line source, particularly in the horizontal direction (along the x-axis). This differential effect appears to be mitigated by both increasing effective iterations and increasing filter widths.

Figure 8.3 compares the effects of increasing effective iterations and altering the reconstruction method. For brevity, only images produced using a 0mm Gaussian filter are shown.

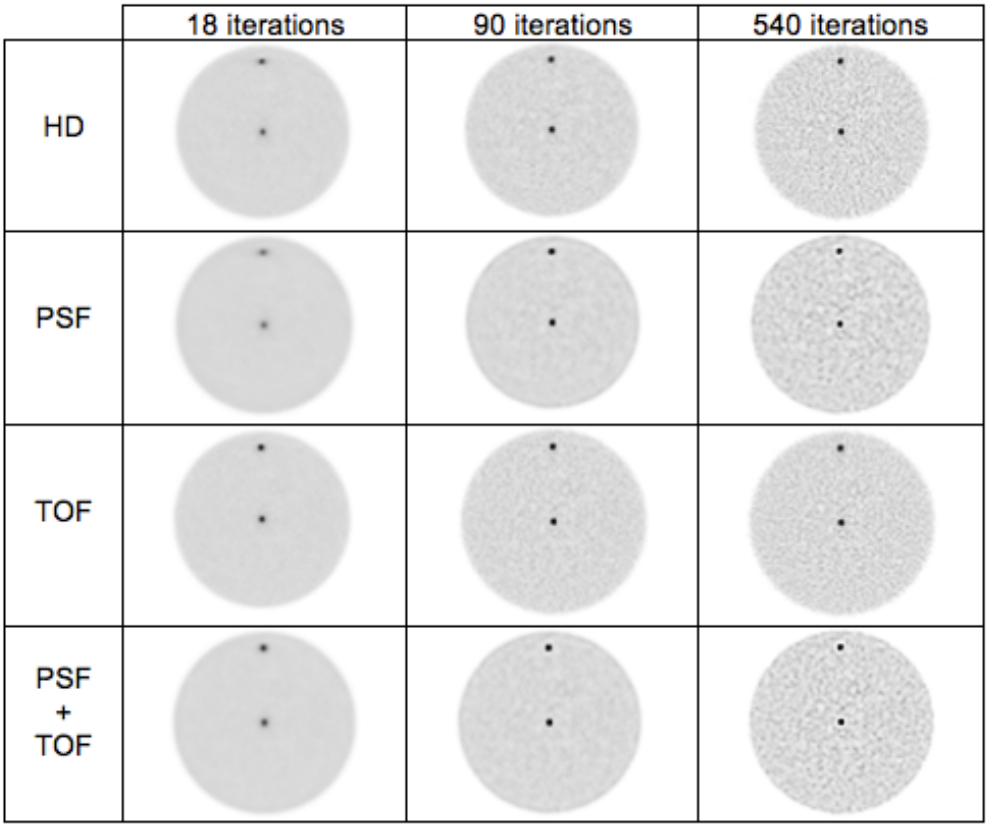


Figure 8.3: Effects of varying reconstruction method and effective iterations 0mm Gaussian filter used for each image.

Figure 8.3 demonstrates that spatial resolution is dependent upon the application of PSF and/or TOF. When TOF-only is applied with 18 iterations, both line sources appear to be sharper and more intense than the HD reconstruction with 18 effective iterations. TOF also appears to produce reasonably consistent resolution for both the central and peripheral line sources, even at low iterations, unlike the HD reconstruction. TOF is known to require less effective iterations to reach convergence, so it is expected to improve image quality at lower iterations. However, TOF is not traditionally reported as directly improving spatial resolution [131], [201].

When PSF-only is applied with 18 iterations, spatial resolution appears to be degraded in comparison with the HD reconstruction, particularly for the peripheral line source along the x-axis. This is not unexpected, as PSF is known to require more effective iterations to reach convergence (as discussed earlier in this thesis). As effective iterations are increased, line sources appear to be sharper and more intense as a result of PSF. There may be a subtle ringing artefact around the peripheral line source when 90 iterations are applied. Ringing artefacts are present around both line sources when 540 iterations are applied but appear more pronounced around the peripheral line source.

PSF+TOF combines the effects of both correction methods. At low iterations, the line sources are not as sharp as those produced by TOF-only, but the resolution of both the central and peripheral line sources appear to be reasonably consistent, unlike PSF-only. At high iterations, ringing artefacts are observed around both line sources, caused by PSF.

8.4.2 Quantitative Analysis of Phantom Images

For brevity and clarity, the effects of Gaussian filter width, effective iterations and reconstruction method will be assessed using only the horizontal profile (x-axis) from the central line source in the first instance. The effects of profile direction (x or y axis) and the line source's position within the FOV will then be assessed.

8.4.2.1 Effects of Gaussian Filter, Effective Iterations, PSF and TOF

Figure 8.4 plots the mean FWHM results for the central source horizontal (x-axis) line profile against effective iterations for each reconstruction method. Results are split into separate plots for each Gaussian filter width setting, with the red box denoting the GEMS suggested clinical filter width (4mm). Figure 8.4 therefore demonstrates the combined effects of the Gaussian filter width, effective iterations and reconstruction method. The dashed red lines in each plot denote the 54 effective iterations suggested by GEMS for clinical imaging.

Figure 8.4 demonstrates that FWHM increases as filter width increases, as one would expect. There is very little difference between the 0mm and 2mm results. Once the filter width is increased beyond 2mm, there appears to be a linear relationship between filter width and FWHM, for all iteration settings and reconstruction methods. Chapter 6 demonstrated that, when a 2.73mm voxel size is used, the Gaussian filter only began to affect image noise when its width exceeded approximately 2mm (see Figure 6.8). Figure 8.4 is therefore consistent with the previous chapter's findings. Figure 8.4 also demonstrates that FWHM decreases as the number of iterations increases in each plot. This is also as one would expect, as increasing effective iterations is known to improve

resolution (and was demonstrated in the previous chapter); however, only minimal FWHM improvements are achieved beyond 180 effective iterations.

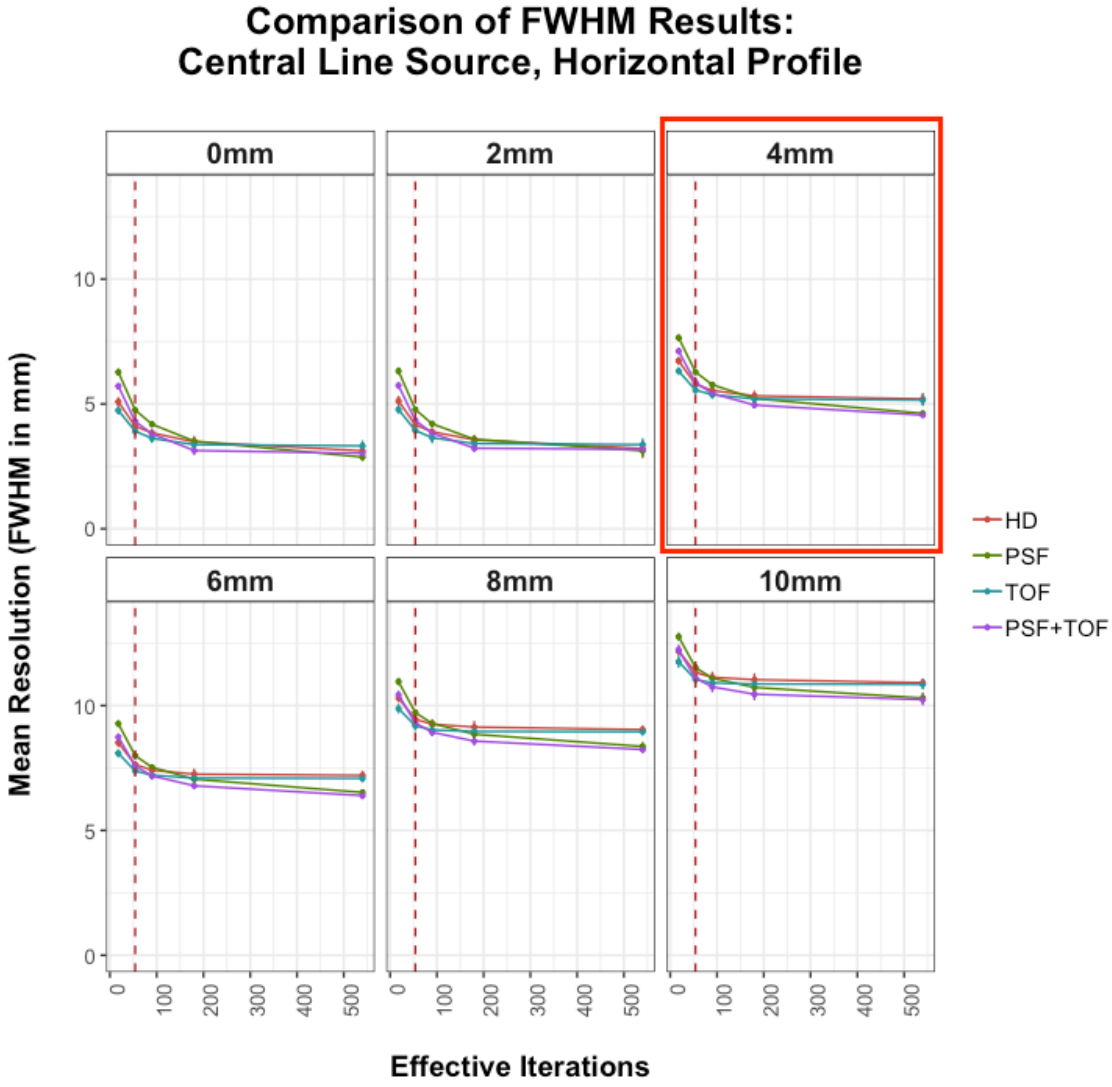


Figure 8.4: Spatial resolution measurements for central line source

Measurements made using a horizontal profile. Data points represent average result over 9 slices. Error bars denote 95% confidence intervals. Dashed red lines and the red box denote the 54 effective iterations and 4mm Gaussian filter suggested by GEMS for clinical reconstructions.

Figure 8.4 also demonstrates that the relative performance of each of the reconstruction methods is dependent upon both effective iterations and Gaussian filter width. At 18 effective iterations, the relative performance of all 4 reconstruction methods remains the same for all filter widths: TOF produces the smallest FWHM values (4.7mm for 0mm filter), followed by HD (5.1mm for 0mm filter) and then PSF+TOF (5.7mm for 0mm filter), while PSF produces the largest FWHM values (6.3mm for 0mm filter). These results are consistent with the qualitative analysis of Figure 8.3, which showed the TOF reconstruction

produced the sharpest images, while the PSF reconstruction produced the blurriest images, when only 18 effective iterations were applied.

When 54 effective iterations are applied, as suggested by GEMS, the relative performance of the reconstruction methods is dependent upon filter width. When the filter width ≤ 4mm, the relative performance remains the same as that of 18 iterations. However, when the filter width exceeds 4mm, the relative performance of PSF+TOF and HD are swapped: TOF FWHM values remain the smallest, followed by PSF+TOF and then HD, with PSF remaining the largest. Although PSF-only continued to produce the poorest resolution at 54 iterations, the improvement PSF+TOF relative to HD-only indicates the increase in iterations, combined with increased filtering, resulted in some relative PSF improvement (6mm filter FWHM results: HD = 7.7mm, PSF = 8.0mm, TOF = 7.4mm, PSF+TOF = 7.6mm).

All 4 reconstructions become more similar at 90 effective iterations, particularly at larger Gaussian filter widths. However, the results still demonstrate some relative differences. When the filter width ≤ 4mm, TOF remains the smallest, followed by PSF+TOF and then HD, with PSF remaining the largest. Once the filter FWHM ≥ 6mm, the relative performance of TOF and PSF+TOF are swapped; PSF+TOF reconstruction now produces the smallest values. When the filter width = 10mm, the relative performances of PSF and HD are swapped: HD now produces the largest values. Both PSF reconstructions therefore demonstrated further improvements as a result of increased iterations and filtering (10mm filter FWHM results: HD = 11.1mm, PSF = 11.0mm, TOF = 10.9mm, PSF+TOF = 10.7mm).

When 180 effective iterations are applied with a filter ≤ 2mm, PSF+TOF produces the smallest FWHM values, followed by TOF, then HD, with PSF producing the largest values. When a 4mm filter is applied, the HD reconstruction produces the largest values, swapping its relative position with PSF. When the filter is ≥ 6mm, the PSF produces smaller values than both HD and TOF, with PSF+TOF producing the smallest values. Again, this illustrates the progressive improvement of PSF as a result of increasing iterations (6mm filter FWHM results: HD = 7.3mm, PSF = 7.0mm, TOF = 7.1mm, PSF+TOF = 6.8mm).

Finally, when 540 effective iterations are applied with a filter width of at least 4mm, one can observe a clear divergence in performance between the reconstructions that use PSF and the reconstructions that do not use PSF. HD and TOF appear to have reached convergence after 180 iterations; extra iterations do not result in further spatial resolution improvements. In contrast, PSF and PSF+TOF demonstrate some further improvements at 540 effective iterations and may improve further with additional iterations. However, these improvements are relatively minor compared to the improvements achieved by increments at lower iterations; any spatial resolution improvements achieved by applying more than 180

effective iterations are unlikely to be clinically significant. The 4mm filter results were as follows: HD = 5.2mm, PSF = 4.6mm, TOF = 5.2mm, PSF+TOF = 4.6mm.

The observations made from the horizontal profile through the central source were mirrored by the other three result sets (vertical profile through central source; horizontal profile through peripheral source; vertical profile through peripheral source).

8.4.2.2 Consistency of Central and Peripheral Resolution

Figure 8.5 compares central and peripheral FWHM measurements made using horizontal profiles. Results are represented as line plots drawn between pairs of results for both line source positions. Non-zero line gradients indicate FWHM discrepancies between the two positions. Each plot within Figure 8.5 compares the measurements made for a particular combination of Gaussian filter width and effective iterations; only three filter widths and three iteration settings are shown for brevity but are sufficient to demonstrate result trends.

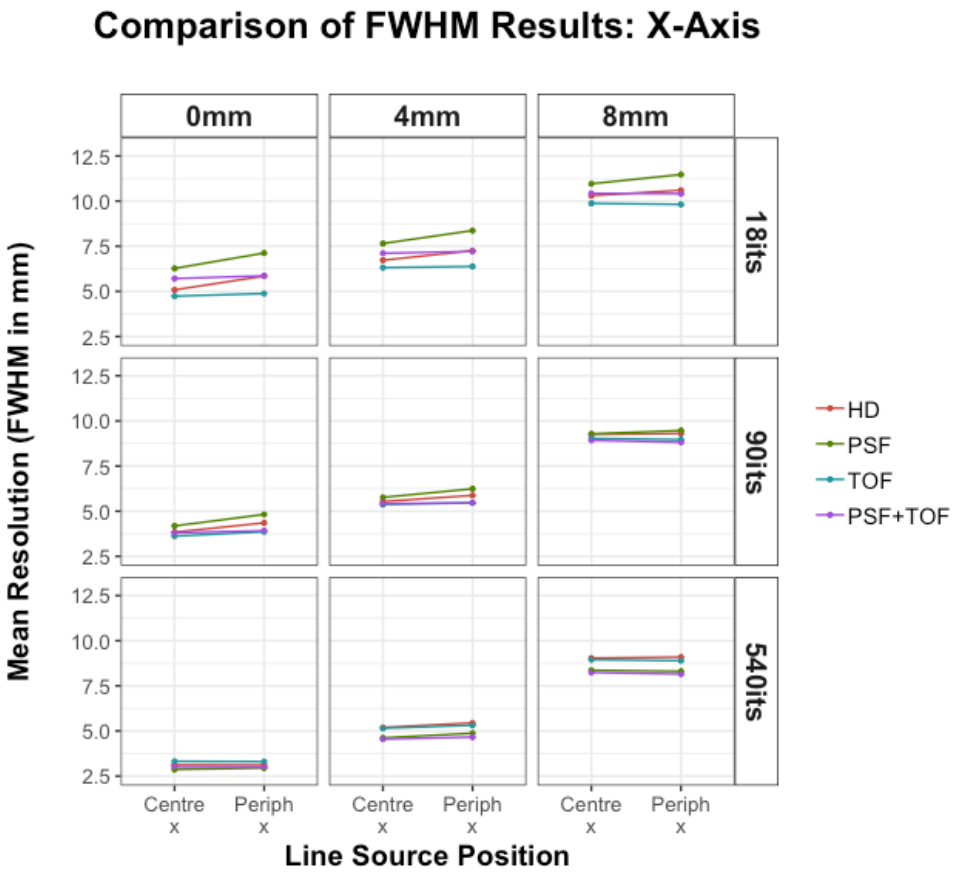


Figure 8.5: Spatial resolution measurements for both line sources, measured using horizontal profiles

Data points represent average result over 9 slices (error bars omitted for clarity).

The HD reconstruction demonstrates poorer peripheral resolution along the x-axis when low iterations are combined with low filtering, as demonstrated by the non-zero line

gradients (HD FWHM discrepancy is 0.76mm when 18 iterations and 0mm filter applied). HD discrepancies reduce as both filtering and iterations are increased, falling to 0.02mm when 540 iterations are combined with a 10mm filter. PSF-only demonstrates similar FWHM worsening with distance from the FOV centre as HD, as demonstrated by their similar, non-zero line gradients (PSF discrepancy is 0.85mm when 18 iterations and 0mm filter applied). As with the HD results, PSF discrepancies reduce as both filtering and iterations are increased, falling to 0.13mm when 540 iterations are combined with a 10mm filter.

In contrast, both TOF reconstructions produce reasonably consistent FWHM results at all iterations and filter combinations, as indicated by the near-zero line gradients. When 18 iterations are combined with 0mm filter, FWHM discrepancies are 0.14mm (TOF) and 0.15mm (PSF+TOF); approximately 20% of the equivalent HD and PSF discrepancies.

Figure 8.6 compares central and peripheral FWHM measurements made using vertical profiles.

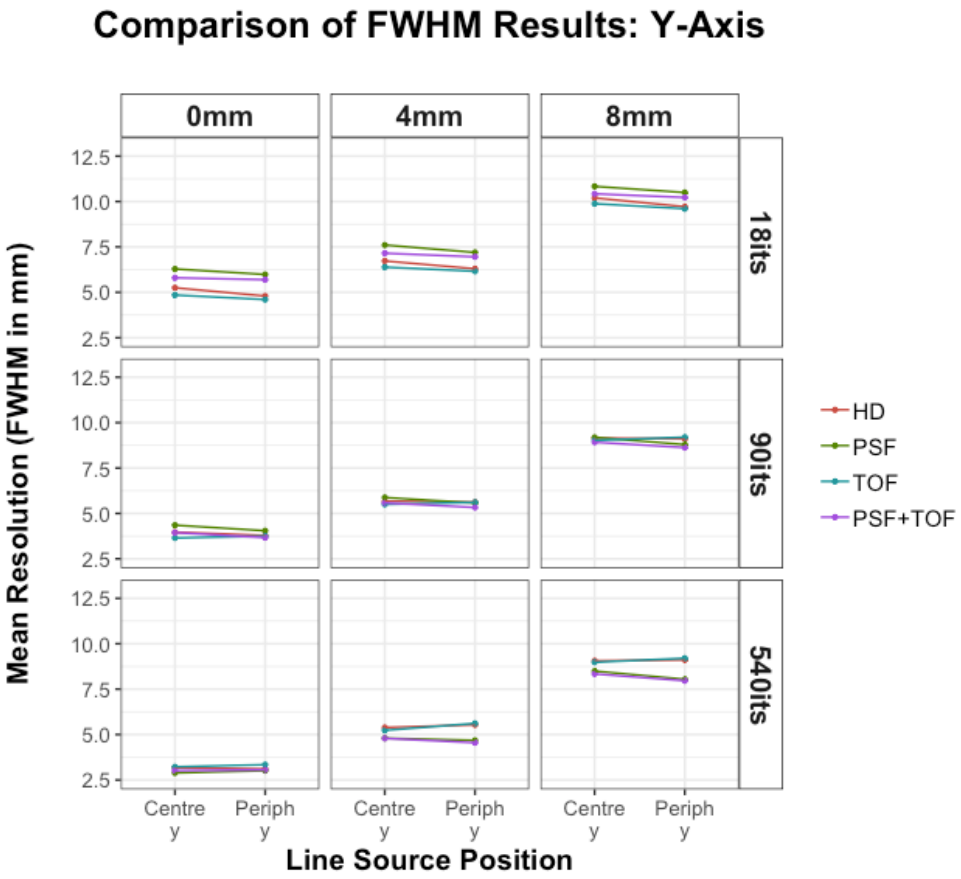


Figure 8.6: Spatial resolution measurements for both line sources, measured using vertical profiles

Data points represent average result over 9 slices (error bars omitted for clarity).

Central and peripheral FWHM results along the y-axis are generally more consistent than those of the x-axis. However, the peripheral line source achieves lower FWHMs than the

central line source for many of the reconstructions used; this is in contrast to the x-axis results.

HD demonstrates marginally poorer central resolution along the y-axis when low iterations are applied, as demonstrated by the non-zero line gradients (HD discrepancy is 0.45mm when 18 iterations and 0mm filter applied). As with the x-axis results, FWHM discrepancies reduce as both filtering and iterations are increased, falling to 0.22mm when 540 iterations are combined with a 10mm filter.

TOF-only also demonstrates marginally poorer central resolution along the y-axis when low iterations are applied (TOF discrepancy is 0.25mm when 18 iterations and 0mm filter applied). TOF discrepancies also reduce as both filtering and iterations are increased, falling to 0.16mm when 540 iterations are combined with a 10mm filter. TOF results are generally more consistent than those of HD for all combinations of filters and iterations.

PSF-only demonstrates similar patterns to HD at lower iterations as demonstrated by their similar, non-zero line gradients (PSF FWHM discrepancy is 0.30mm when 18 iterations and 0mm filter applied). However, the PSF discrepancies remain apparent after 540 iterations, particularly when combined with filtering (PSF FWHM discrepancy is 0.58mm when 540 iterations and 10mm filter applied).

PSF+TOF FOV discrepancies mirror those of TOF at lower iterations (discrepancy is 0.10mm when 18 iterations applied with 0mm filter), and those of PSF at higher iterations (discrepancy is 0.38mm when 540 iterations applied with 10mm filter).

These quantitative results are consistent with the qualitative analysis of Figure 8.3:

- PSF appeared to cause increased peripheral source blurring along the x-axis at low iterations, confirmed by the 0.85mm discrepancy in x-axis FWHMs when 18 iterations were applied (compared with 0.76mm for HD)
- TOF-only appeared to produce the most consistent central/peripheral x-axis FWHMs at low iterations, confirmed by the 0.14mm discrepancy in x-axis FWHMs when 18 iterations were applied.
- Peripheral x-axis blurring was reduced by increasing iterations and was not visualised after 540 iterations for any reconstruction (HD and PSF discrepancies fell to 0.02mm and 0.13mm respectively when a 10mm filter was applied)

FWHM discrepancies along the y-axis were difficult to visualise in Figure 8.3, which only included reconstructions produced with a 0mm filter. Figure 8.6 confirms that FWHM

discrepancies along the y-axis were minimal at all iterations when no filter was applied. Larger discrepancies were demonstrated along the y-axis for the PSF and PSF+TOF reconstructions when 540 iterations were combined with wider Gaussian filters (0.58mm and 0.38mm respectively when a 10mm filter was applied); however, these are equivalent to approximately 5% discrepancies between central and peripheral FWHMs and are therefore unlikely to be clinically significant.

8.5 Discussion

Previous studies that assessed spatial resolution using FBP point/line sources in air demonstrated that PSF significantly improved spatial resolution, and equalised resolution throughout the FOV, when compared with non-PSF reconstructions. However, the absence of surrounding activity makes these experimental scenarios clinically unrealistic: they cannot facilitate a valid assessment of iterative reconstruction techniques (and of PSF in particular). This was demonstrated by the Alessio paper [49], which compared measurements of line sources in both air and water for a Discovery STE PET-CT system (which had BGO crystals and therefore no TOF capability). This chapter aimed to assess the relative effects of reconstruction parameters upon spatial resolution in a clinically relevant manner for the Discovery 690 PET-CT system, which has both PSF and TOF capabilities. Line sources within background radioactivity were acquired and reconstructed using 120 combinations of reconstruction parameters (effective iterations, post-reconstruction filter widths, TOF and PSF) and FWHM measurements from the resultant images were used to assess the effects of the reconstruction parameters.

Both qualitative and quantitative analysis demonstrated that effective iterations and post reconstruction filtering impacted spatial resolution in an expected manner: resolution improved when more effective iterations were applied and degraded when the extent of post-reconstruction filtering was increased.

TOF was shown to improve resolution of both the central and peripheral line sources, and improved the consistency between them, when compared to HD after only 18 iterations. TOF is known to reach convergence after fewer iterations than plain OSEM [211] so an early improvement in image quality is not unexpected. However, it has been published that TOF “does not improve spatial resolution” [19], so these low-iteration TOF findings may, at first glance, appear to be unexpected. Figure 8.4 plotted FWHM versus effective iterations for various filter widths: for each filter width, TOF appeared to be the first of the four reconstruction methods to converge to a final resolution measurement, and there was minimal FWHM improvements once 90 effective iterations are exceeded. TOF converged

to similar values as the HD reconstruction, so it is not unreasonable to state “TOF does not improve spatial resolution”; however, it does converge faster than HD, which means at low matched iterations it appears to improve spatial resolution.

When low numbers of effective iterations were used, PSF degraded spatial resolution throughout the image, and worsened FWHM discrepancies along the x-axis between central and peripheral line sources when compared to HD. PSF is known to require more iterations to reach convergence, as the system matrix incorporates more information into the reconstruction algorithm [89], so it is not unexpected PSF image quality is poorer than that of HD at low iterations. Increased peripheral horizontal blurring as a result of using PSF with low iterations is also consistent with that observed in Alessio’s study [49]. However, when effective iterations were increased beyond approximately 150, PSF produced better spatial resolution than HD, and produced reasonably consistent FWHM results at both FOV positions. Figure 8.4 also demonstrated PSF does not appear to reach convergence after 540 effective iterations, indicating that further improvements in spatial resolution could be achieved if more iterations were applied.

Applying both PSF and TOF together was shown to combine the characteristics of the TOF and PSF, as one would expect. While resolution remained poorer than HD after 18 iterations as a result of PSF, the central and peripheral line sources produced reasonably consistent FWHMs, as a result of TOF. PSF+TOF produced superior resolution measurements when over 90 iterations were applied, and appeared to converge to lower FWHM values than either HD or TOF-only. Furthermore, PSF+TOF converges faster than PSF-only.

Although they did not perform quantitative spatial resolution assessments, Bettinardi et al [19] found that the best overall image quality results on the GEMS Discovery 690 were obtained by applying both TOF and PSF together, stating “…if TOF acts as an accelerator for signal convergence, PSF can recover a better signal at a lower number of iterations…”. The results of this chapter appear to indicate this is also the case for when considering quantitative spatial resolution measurements, provided enough effective iterations are applied. However, the final choice of reconstruction parameters will not be solely dependent upon spatial resolution, as discussed elsewhere in this thesis.

Finally, the purpose of PSF modelling is often described in the literature as ‘to improve spatial resolution’ [19], [137], [138], [146], [212]. The results of this chapter demonstrate that this is not necessarily the case; at low iterations, PSF was shown to degrade resolution. Any statement regarding the use of PSF to improve spatial resolution must account for effective iterations: up to 180 iterations may be required for this to be true. This is significantly more effective iterations than used by papers that concluded PSF improved

resolution after assessing sources in air (Murata et al [197] used up to 54 effective iterations; Suljic et al [201] used 84 effective iterations). One must conclude these studies would have drawn different conclusions had their sources been assessed under more clinically realistic conditions. It is of interest the GEMS' White Paper describes the purpose of their 'SharpIR' algorithm as "…to improve PET image contrast to noise…" [89] rather than explicitly stating it improves spatial resolution. The work in this chapter suggests that, although PSF can improve spatial resolution when enough effective iterations are applied, the number of required effective iterations are higher than typically used clinically. It may therefore be misleading to describe the main benefit of PSF modelling as "improving spatial resolution".

A limitation of this study was that a single phantom size and only two FOV positions were assessed. Future work could involve a larger phantom containing several line sources at different positions throughout the FOV. The use of perpendicular line source positioning within the phantom would allow FWHM measurements along the x, y and z axes from a single acquisition.

8.6 Conclusions

This chapter demonstrated that the use of PSF with very high effective iterations (>540) and no filtering produced the optimal clinical resolution measurements. However, such high iterations are unlikely to be used clinically due to unacceptable image noise, as previously discussed in Chapter 6.

This chapter also demonstrated that, contrary to what has been reported in some publications, PSF-only reconstructions degrade clinical spatial resolution at lower iterations (<180), which are more likely to be used clinically. Therefore, the use of either TOF-only (up to approximately 90 iterations), or PSF and TOF together (between approximately 90 and 180 iterations), are more appropriate for maximising spatial resolution.

The results of this chapter will be combined with the assessments of noise and contrast recovery, in order to select the optimal reconstruction parameters for detecting small lesions within the liver. Contrast recovery is investigated in the following chapter.

Chapter 9 : Assessment of Lesion Detection in Reconstructed Images of a Body Phantom

The primary aim of this thesis is to develop a generic methodology to assess and optimise PET image reconstruction whilst considering clinical context. The effects of reconstruction parameters upon noise and spatial resolution were assessed in previous chapters; this chapter builds on that work with a view to optimising the detection of simulated lesions of varying sizes within a body phantom. The use of different image quality metrics will also be assessed, with a view to making recommendations for generic optimisation tasks.

The secondary aim of this thesis is to optimise PET image reconstruction on the GEMS Discovery 690 for the detection of small liver lesions. This chapter therefore uses the body phantom to represent ^{18}F-FDG liver imaging and determine which combinations of reconstruction parameters are likely to optimise small liver lesion detection.

9.1 Introduction

Lesion detectability has been assessed by multiple studies using different methodologies. Mathematical observers have been designed to predict human observer performance in lesion detection tasks when the number of images to be assessed is very large, and therefore beyond what human observers could realistically assess (tens/hundreds of thousands of images) [213]. Such observers have typically been used to calculate lesion SNR in both phantom and patient images [8], [157], [213]–[220]. However, lesion detection performance of human observers is best assessed using human observers, as opposed to mathematical observers [214]; studies using mathematical observers are often validated by subsequent human observer studies of smaller data sets. As mathematical observers are

complex to develop, and the number of reconstructions to be compared in this thesis was small (in relative terms), they were not considered appropriate for this work and are therefore beyond the scope of this thesis. However, the fact that many mathematical observers perform SNR calculations when assessing lesion detection is of interest. SNR reflects the relative signal level with respect to noise in the reconstructed image, and is therefore a widely accepted metric for lesion detection [221], [222]. SNR has also been used in many conventional studies performing quantitative analyses of image quality [18], [111], [134], [158], [223].

The relationship between image noise and contrast recovery has a direct impact on lesion detection [192]. Contrast recovery measurements of simulated hot lesions within background activity, where the radioactivity concentrations of both the simulated lesions and the background are known exactly, are widely used in phantom studies assessing image quality. Contrast recovery analyses using lesions of varying sizes can also demonstrate the impact of partial volume effects, caused by limitations in PET spatial resolution (as discussed in Chapter 1) [77], [159].

Dual-metric trade-off curve analysis, e.g. plotting noise against lesion metrics such as contrast recovery [49], [90], [130], [133], [169], [189], [224]–[227], SNR [92], resolution [15], [16] or contrast [18], [134], [145], have been described as a useful first step when comparing different reconstruction techniques; they can be used to narrow down the range of parameters to be used in subsequent observer studies [37].

Several publications have assessed hot sphere contrast recovery by following NEMA's image quality assessment methodology, using the NEMA IEC Body phantom (described in Section 3.2.2). This methodology also involves measuring Background Variability (described in Section 3.4.2) using ROIs size-matched to each sphere. NEMA image quality assessment results are typically compared with manufacturer specifications; however, many studies have extended the methodology to assess the effects of reconstruction parameters, incorporating other image quality metrics such as SNR, SUVs etc.

A 2011 study by Bettinardi et al [19] used the NEMA body phantom and methodology to assess contrast recovery produced by a GEMS Discovery 690 PET-CT system. They assessed four different reconstruction methods (HD, PSF, TOF and PSF+TOF) as effective iterations were increased from 18 to 360 (the effects of altering filters or voxel sizes were not assessed). Background COV was used as a noise metric. Hot contrast recovery coefficients (HCRCs, defined by NEMA) were plotted against effective iterations and against COV. The study concluded that applying PSF and TOF together achieved the best image quality results. They did not make a specific recommendation for effective iterations,

only stating that approximately 54 effective iterations are used clinically to restrict reconstruction times. They also stated the differences in noise characteristics caused by TOF and PSF could play an important role in terms of lesion detectability and requires further work. SNR was not assessed in this study.

Jakoby et al [111] focussed on the effects of adding TOF to PSF reconstructions, also performing NEMA style lesion detection assessments on a Siemens Biograph PET-CT system. The effects of adjusting filters were not assessed by this study. PSF-only and PSF+TOF reconstructions used 63 and 42 effective iterations respectively; these produced similar background variation results, with TOF producing greater sphere contrast results. This study included SNR measurements for patient reconstructions, which used Standard Deviation of a large background region (i.e. a voxel-to-voxel noise measurement). TOF was shown to improve SNR by a factor of 1.2 for a small patient (BMI = 21.6) and by a factor of 1.7 for a larger patient (BMI = 31.6).

Lois et al [18] also used the NEMA phantom to assess the impact of adding TOF to PSF reconstructions for lesion detectability. TOF maximised SNR after 28 effective iterations when compared to 56 effective iterations for non-TOF. These quantitative results were confirmed by qualitative analysis. No other reconstruction parameters were assessed by this study.

Kuhnert et al [149] used the NEMA phantom to measure the effect TOF and PSF had upon SUV_{max} measurements using a Siemens Biograph mCT PET-CT system. This study only compared two reconstruction methods: the first reconstruction applied both PSF and TOF with 63 effective iterations and a 2mm FWHM Gaussian filter (following the manufacturer's recommendations), and the second reconstruction applied OSEM without TOF or PSF, using 48 effective iterations and a 5mm FWHM Gaussian filter (following clinical trial certification criteria). The study found statistically significant differences in SUV_{max} measurements between both reconstructions, particularly for the smaller sphere: the PSF+TOF reconstruction SUV_{max} results approximately doubled for the smallest two spheres, due to the known Gibbs artefacts caused by PSF. This improved small sphere detectability but compromised quantitative accuracy. The study recommended the use of two data sets for image interpretation: one set to optimise visual assessment (using TOF and PSF) and one set for standardised quantitative image interpretation (e.g. without PSF, or PSF applied with extra filtering).

Schaefferkoetter et al [8] added artificial liver lesions to 40 sets of patient images and assessed the effects of PSF, TOF, effective iterations and Gaussian filtering. A Siemens Biograph PET-CT system was used in this study. Lesion SNR was primarily assessed using

a mathematical observer, with a restricted reconstruction set assessed by human observers. They concluded lesion detection was optimised when both PSF and TOF were applied together, with 24 effective iterations and no Gaussian filter applied. Other studies using mathematical observers to assess PSF and TOF, whilst not making specific recommendations for effective iterations or Gaussian filtering, also concluded that applying both PSF and TOF together maximised lesion SNR, and therefore maximised lesion detection [157], [215], [217], [220].

This literature review demonstrates that the optimal reconstruction parameters for the GEMS Discovery 690 have not been established; in particular, no recommendations have been made to optimise liver lesion detection on the GEMS Discovery 690. Furthermore, SNR analysis in the literature appears to be exclusively based upon voxel-to-voxel measurements; to the authors knowledge, the use of SNR based on region-to-region noise measurements has not been assessed.

9.2 Aims

This chapter aims to characterise the effects of the following reconstruction parameters upon clinical lesion detection within a body phantom, with a view to identifying the combination of parameters that maximise small sphere detection:

a. Reconstruction method (TOF and PSF)
b. Effective iterations
c. Gaussian filter width

Lesion detection will be assessed qualitatively and quantitatively. As previously discussed, the relationship between image noise and contrast recovery has a direct impact on lesion detection in PET images [192]. HCRC and noise measurements (both voxel-to-voxel Image Roughness and region-to-region Background Variation) will therefore be assessed. HCRC measurements will also provide an indication of spatial resolution: partial volume effects will be more apparent as the spheres decrease in size. Finally, SNR will be assessed for the spheres. A novel method of SNR calculation will also be assessed: Background Variation will be used to represent image noise.

The work in this chapter will be performed in three parts:

- **Part 1:** a single phantom acquisition will be used to assess a similar, but slightly reduced, range of reconstruction parameters evaluated by previous chapters (as justified in the Materials and Methods section of this chapter). The aim of Part 1 is to further narrow down the range of relevant values for each reconstruction parameter and

exclude parameter combinations that are not appropriate for maximising small lesion detection.

- **Part 2:** A narrowed-down range of values for each reconstruction parameter, using finer increments where appropriate, will be assessed for the same single phantom acquisition used in Part 1. The aim of Part 2 is to determine the optimum combination of reconstruction parameters for small sphere detection.
- **Part 3:** The reconstructions used in Part 2 will be applied to multiple phantom acquisitions to assess the reproducibility of this chapter's methodology.

This chapter will conclude by recommending a combination of reconstruction parameters to be used for detecting small lesions within the liver.

9.3 Materials and Methods

9.3.1 Body Phantom and Acquisition Protocol

The NEMA IEC body phantom (described in Section 3.2.2) was used in this chapter. A sphere-to-background ratio of approximately 4:1 was used as this is suggested by the NEMA guidelines for assessing image quality [177]. Table 9.1 summarises the phantom activity concentration at the time of acquisition and the sphere-to-background ratio used in Parts 1 and 2 of this chapter. The activity concentration in the background chamber was chosen to be similar to that found within the livers of patients undergoing ^{18}F-FDG imaging (discussed in Chapter 4). Activity concentrations used for Part 3 are detailed later in this chapter.

Background Activity Concentration	Sphere Activity Concentration	Ratio
5,346Bq/ml	24,830Bq/ml	4.64:1

Table 9.1: Body phantom activity concentrations

Phantom used in Parts 1 and 2 of this chapter. Resultant sphere-to-background ratio is also shown.

The phantom was acquired using two four-minute bed positions with a 23% overlap (as justified by Chapter 4). Spheres were positioned within the overlap region in order to assess lesion detection in the area of lowest sensitivity (and therefore the most challenging area for lesion detection, as described in Chapter 4).

9.3.2 Image Reconstruction Protocols

Reconstruction parameters used in Part 1 of this chapter were partly informed by the findings of previous chapters. As one would expect, Chapter 6 demonstrated applying low iterations with heavy filters and large voxels minimised image noise, while Chapters 7 and

8 demonstrated applying very high numbers of iterations in combination with minimal filtering and small voxel sizes optimised spatial resolution. It was therefore not possible to choose optimum voxel sizes, filter widths or effective iterations based only on these conflicting requirements. However, as the secondary aim of this thesis is to optimise the detection of small liver lesions, this chapter uses only the 256 matrix combined with the full 700mm transaxial FOV. This will ensure results are applicable to larger patients, whilst minimising partial volume effects between liver lesions and the surrounding tissue.

The axial voxel size (z-axis direction) for the GEMS Discovery 690 is fixed at 3.34mm, which is relatively large in comparison to the chosen transaxial voxel size (2.73mm). Furthermore, Chapter 6 demonstrated the z-axis filter had minimal effects upon image noise when compared to the transaxial Gaussian filter. This chapter therefore does not assess the effects of the z-axis filter upon lesion detection; no z-axis filter is applied.

Both noise and spatial resolution assessments concluded PSF should be applied; however, the spatial resolution chapter also concluded PSF-only should be used to maximise spatial resolution only if ≥ 180 effective iterations were applied (PSF-only was found to degrade resolution at low iterations). This conflicts with the low iteration requirement for optimising image noise. It is therefore difficult to narrow down the approach to reconstruction methodology (PSF and/or TOF) without further assessment. The literature demonstrates that more than 180 effective iterations are unlikely to be used clinically due to the high noise and increased reconstruction times; however, as PSF is known to require higher number of iterations to improve spatial resolution, 540 effective iterations were included in this chapter to fully demonstrate the effects of PSF upon contrast recovery.

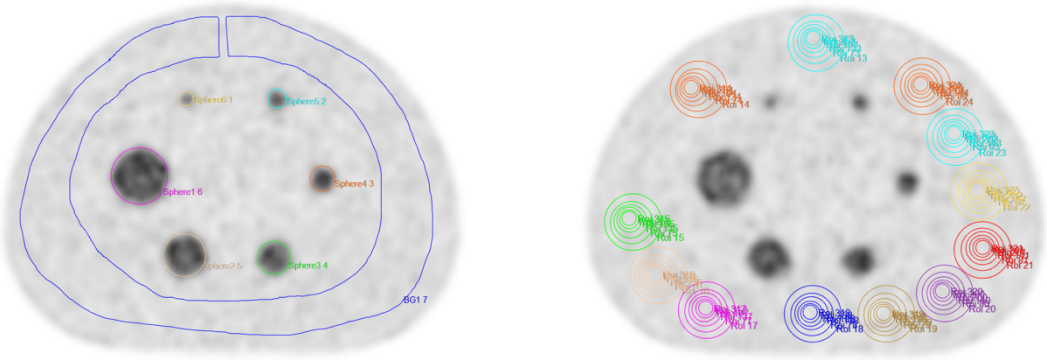
Table 9.2 summarises the reconstruction parameters used in Part 1 of this chapter. Parameters used for Parts 2 and 3 are detailed, and justified, later in this chapter.

Reconstruction Parameter	Settings Used	Number of Result Groups
Reconstruction Method	HD, PSF, TOF, PSF+TOF	4
Effective OSEM Iterations	18, 54, 90, 180, 540	5
Gaussian Filter (FWHM)	0 →10mm, in 1mm increments	11
Z-Axis Filter	None	1
Matrix Size	256*256 only	1

Table 9.2: Reconstruction parameters used for body phantom (Part 1 only)

9.3.3 Image Analysis

All ROI measurements were performed using Hermes Medical Systems' Hybrid Viewer software. Two ROI templates were used: the first one was used for the sphere analysis and the second one was used for background noise analysis, as shown in Figure 9.1.



(a) (b)

Figure 9.1: ROI profiles on PET images for (a) HCRC analysis and (b) noise analysis Slice showing maximum extent of spheres chosen for sphere ROIs. Background ROIs were also drawn on two slices either side of this slice.

The first ROI template (Figure 9.1 (a)) follows the same methodology used in Chapter 4 to calculate contrast recovery coefficients [179]. The second ROI template (Figure 9.1 (b)) is based upon the NEMA image quality assessment [70]: twelve ROIs matching each of the six sphere sizes are drawn on the central slice, and copied to two slices either side (360 background ROIs in total). This enables separate Image Roughness and Background Variation calculations to be made for each sphere; each calculation uses the 60 ROIs which match the sphere's diameter.

ROIs from both templates were then copied to all PET reconstructions and used to extract the following data:

- Mean Activity Concentration for all six spheres
- Maximum Activity Concentration for all six spheres
- Mean Activity Concentration for all background ROIs
- Standard Deviation for all background ROIs

The following image metrics were then calculated using the extracted data:

- COV of the annular background regions (Equation 3.2)
- Image Roughness (six different results obtained, one for each sphere size) (Equation 3.3)
- Background Variation (six different results obtained, one for each sphere size) (Equation 3.4)

- Mean and Maximum HCRC for each sphere (Equation 3.5)
- Mean and Maximum SNR for each sphere (Equation 3.8), using two different noise metrics:
    - Standard deviation of annular ROIs, i.e. voxel-to-voxel noise measurement widely used in the literature for SNR analysis, and
    - Background Variation, i.e. region-to-region noise measurement, using ROIs matched to the sphere size

9.4 Part 1 Results: Lesion Detection Assessment

This section aims to narrow down the range of reconstruction parameters under consideration and exclude parameter combinations that are not appropriate for maximising small lesion detection. Qualitative analysis of phantom images will determine which combination(s) of effective iterations, Gaussian filtering and reconstruction method (PSF and/or TOF) are preferred by experienced observers when attempting to visualise the smallest sphere in the phantom. These results will then be compared with quantitative analysis to determine which image metrics best correlate with the qualitative observer study.

9.4.1 Qualitative Analysis of Phantom Images

Figures 9.2 to 9.5 show reconstructed phantom images from Part 1 of this chapter. Figure 9.2 shows the HD images, followed by the PSF-only images (Figure 9.3), TOF-only images (Figure 9.4) and finally PSF+TOF images (Figure 9.5). Each figure illustrates the effects of increasing effective iterations and Gaussian filter width (for brevity, not all Gaussian filter results are shown in these figures). All four figures demonstrate that as the filter width increases, image noise decreases and spatial resolution degrades, and that as the number of effective iterations increases, image noise increases and spatial resolution improves. These findings are in accordance with the results of previous chapters and are as one would expect. Additionally, all four figures demonstrate that any differences between the 0mm and 2mm Gaussian filter images are difficult to visualise. This is in agreement with the findings of Chapter 6: when 2.73mm voxels are used, the Gaussian filter width had no effect until it is increased to at least approximately 2mm. Finally, the background noise appears to be of a random nature in both the HD (Figure 9.2) and TOF-only (Figure 9.4) images, and of a correlated nature in the PSF-only (Figure 9.3) and PSF+TOF (Figure 9.5) images. This is consistent with the results in Chapter 6, which discussed this finding in more detail.

Figures 9.2 to 9.5 were qualitatively assessed by four experienced PET physicists in terms of their ability to visualise the smallest sphere (120 images in total). Each of these observers selected the reconstruction they judged to maximise the smallest sphere's visibility.

Figure 9.2 below shows the HD reconstruction images. Each row shows a different Gaussian filter width, while each column shows a different number of effective iterations.

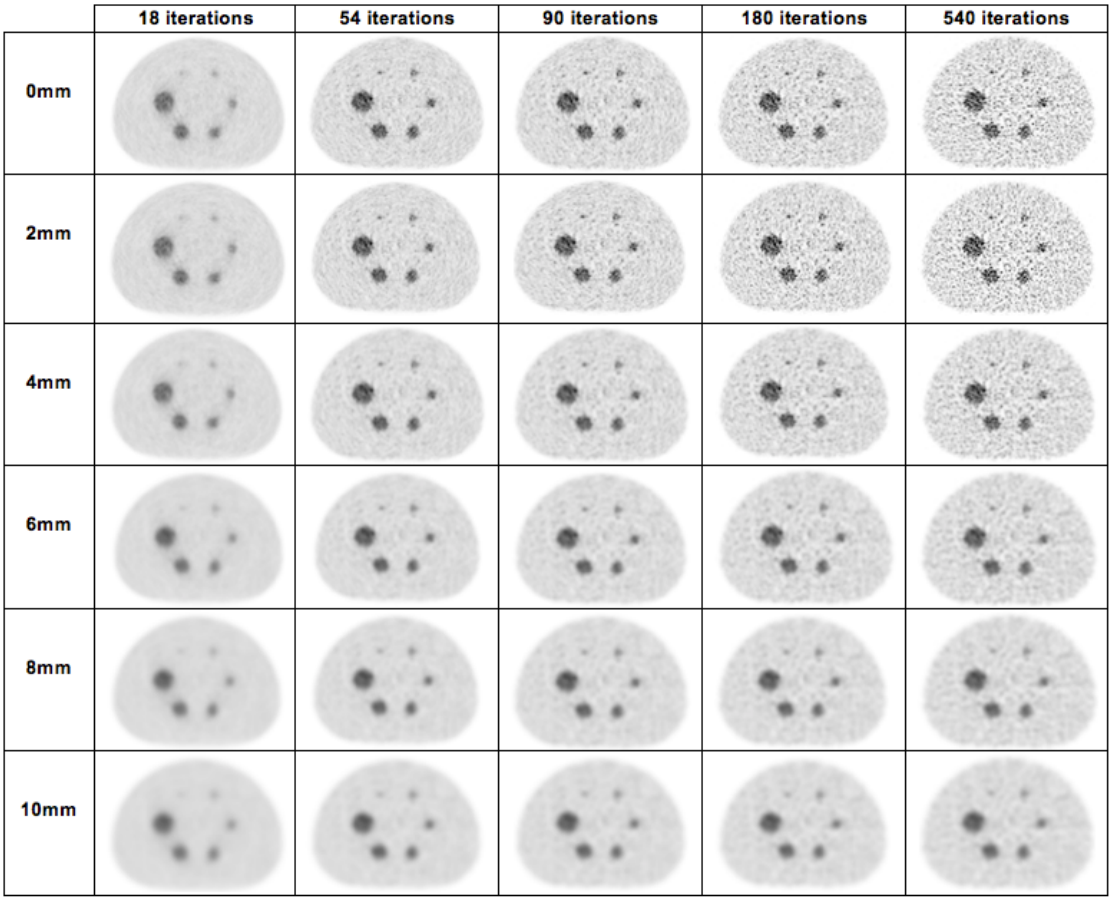


Figure 9.2:HD reconstructed images with varying Gaussian filter widths and effective iterations

The smallest sphere appears to be most easily visualised when between 54 and 90 effective iterations are applied with little or no filtering, when the HD reconstruction method is used: however, none of the HD reconstructions were considered to be optimal by any of the observers.

Figure 9.3 shows the PSF-only reconstruction images.

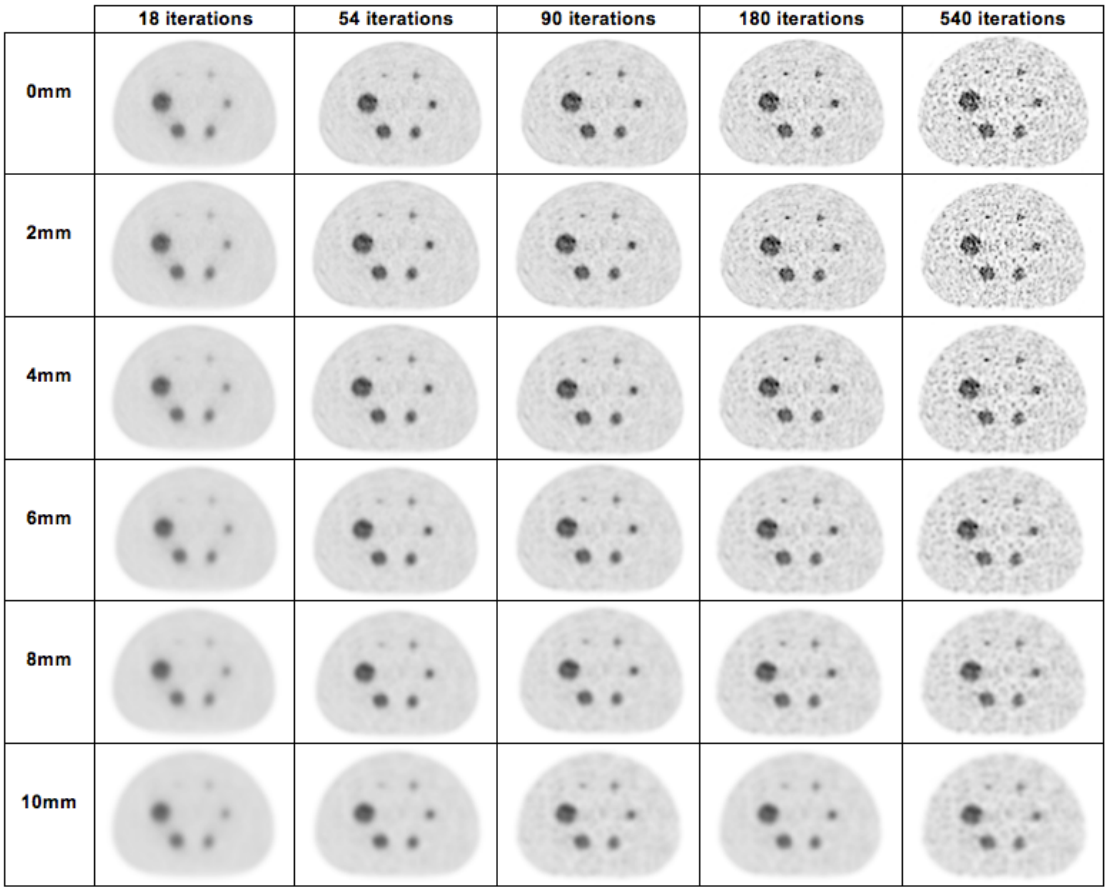


Figure 9.3: PSF-only reconstructed images with varying Gaussian filter widths and effective iterations

The smallest sphere appears to be most easily visualised when between 90 and 180 effective iterations are applied with no filtering, when the PSF-only reconstruction method is used. PSF-only reconstructions appear to offer superior small lesion detection than HD reconstructions; however, PSF-only reconstructions were also not considered to be optimal by any of the observers.

Figure 9.4 shows the TOF-only reconstruction images.

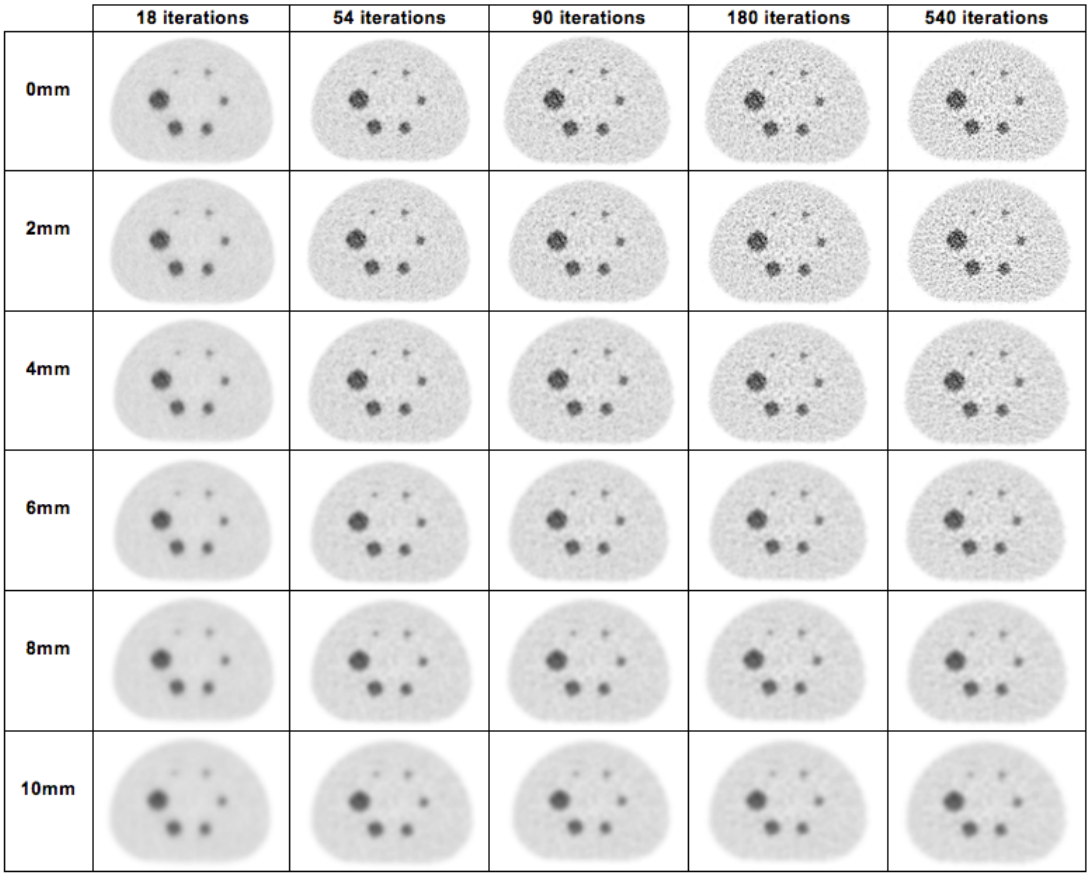


Figure 9.4: TOF-only reconstructed images with varying Gaussian filter width and effective iterations

The smallest sphere appears to be most easily visualised when between 54 and 180 effective iterations are applied with little or no filtering, when the TOF-only reconstruction is used. TOF-only reconstructions appear to offer superior small lesion detection than HD reconstructions, particularly at low iterations; however, it is difficult to determine how its performance compares to that of PSF-only, as the noise characteristics are markedly different. Again, none of the TOF-only reconstructions were considered optimal by any of the observers.

Finally, Figure 9.5 shows the PSF+TOF reconstruction images. All four observers agreed that PSF+TOF image quality was superior to that of the other three reconstruction methods with respect to visibility of the smallest sphere. Three observers preferred the use of 90 effective iterations, with 0mm or 2mm Gaussian filtering. It is of interest that when presented with a total of 120 images, three observers selected the same two images as being optimal. The fourth observer preferred the use of 180 effective iterations, with 0mm or 2mm Gaussian filtering. It is also of interest that no observers selected the reconstruction which most closely followed the GEMS suggestion for clinical reconstructions.

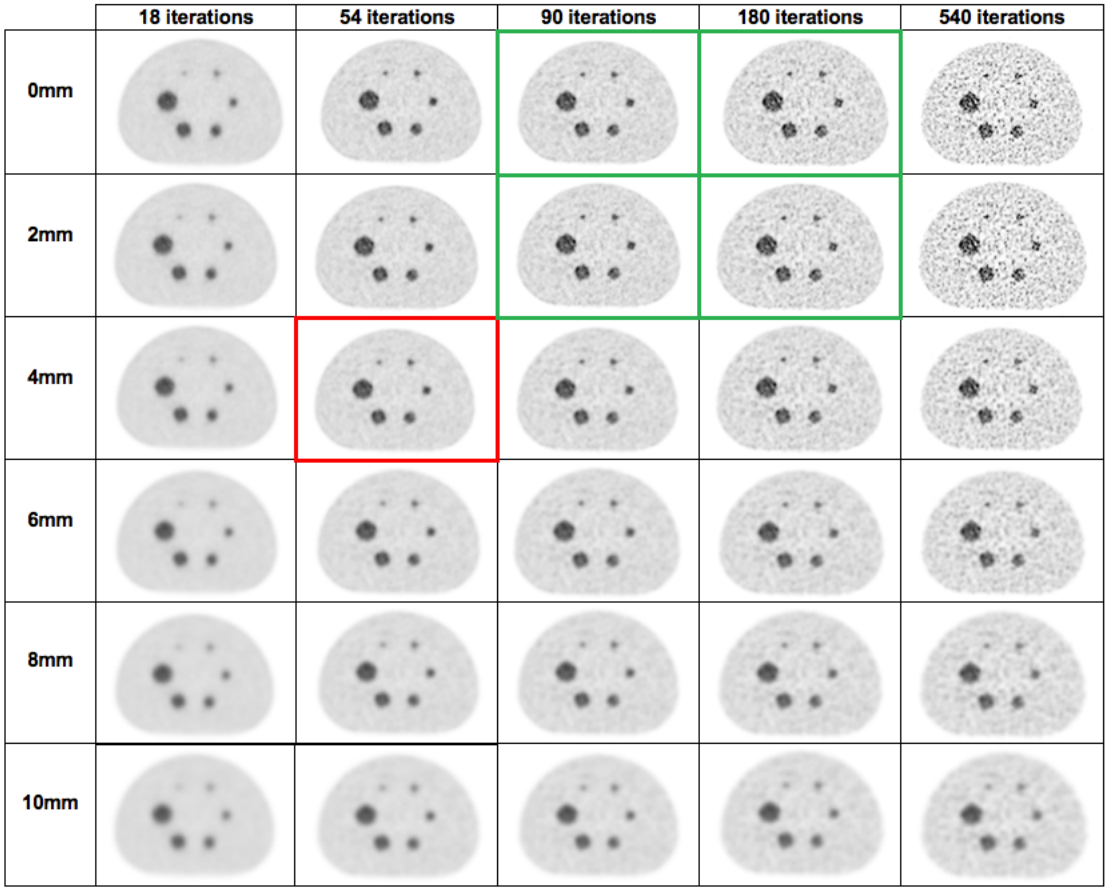


Figure 9.5: TOF+PSF reconstructed images with varying Gaussian filter width and effective iterations

Red box denotes the 54 effective iterations and 4mm Gaussian filter, suggested by GEMS for clinical reconstructions. Green boxes denote reconstructions favoured by observers.

9.4.2 Quantitative Analysis of Phantom Images

This section details the quantitative analysis performed on the phantom images. Firstly, image noise (both Image Roughness and Background Variation) will be discussed briefly, followed by sphere SNR and HCRC as a measure of lesion detection.

9.4.2.1 Noise Results

Voxel-to-voxel noise (Image Roughness) and region-to-region noise (Background Variation) were calculated for all phantom reconstructions undertaken for this chapter. Both of these measurements were previously assessed in Chapter 6; however, only a single ROI size (37mm diameter) was used in that chapter, as the aim was to assess the overall noise within the reconstruction. This chapter aims to relate noise to the detection of small spheres; both Image Roughness and Background Variation are therefore related to each sphere size by using ROIs that match inner sphere diameters, as discussed in Section 9.3.3. This is consistent with the NEMA image quality assessment for Background Variation [70]; however, this technique is not generally used for Image Roughness, or other methods of

voxel-to-voxel noise measurements, as smaller ROIs are known to underestimate voxel-to-voxel noise.

Figure 9.6 plots Image Roughness against effective iterations for all six ROI sizes. Two sets of results are shown: the top row used no Gaussian filter, while the bottom row used a 10mm Gaussian filter.

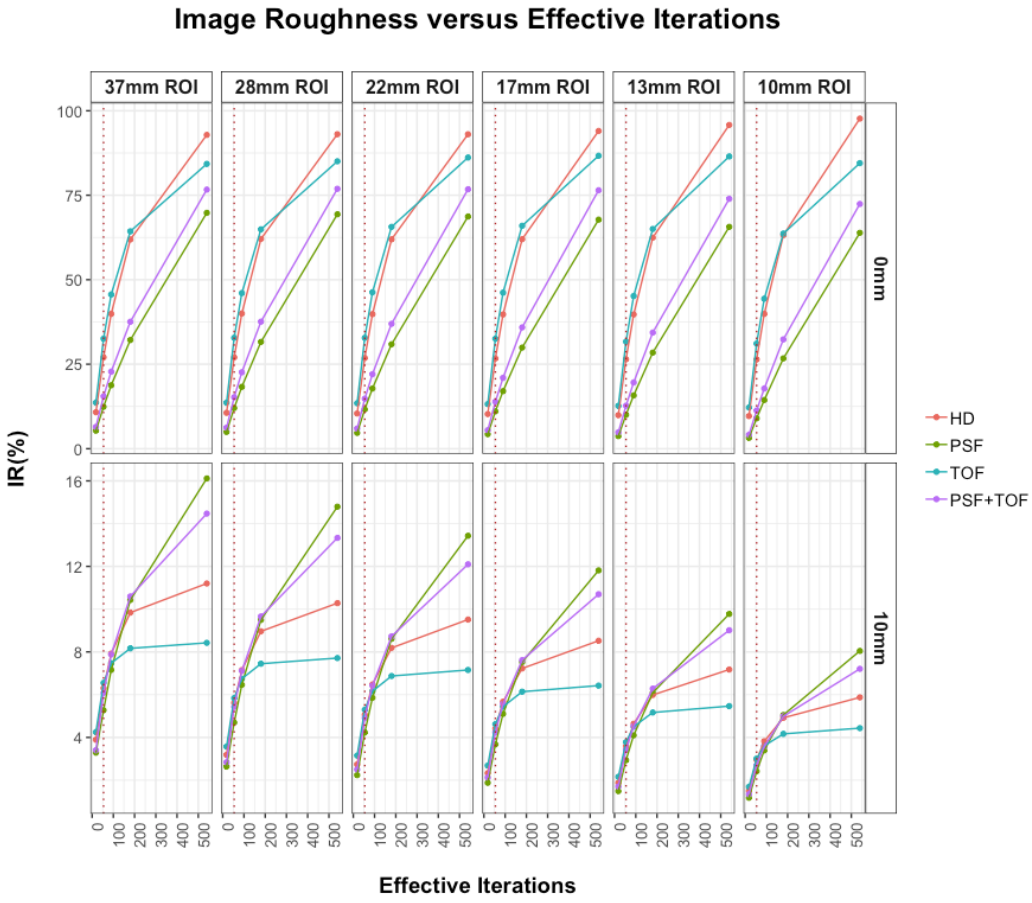


Figure 9.6: Image Roughness (IR) versus effective iterations

Top row uses no Gaussian filter, bottom row uses 10mm Gaussian filter. Results are split by ROI size (which are matched to sphere sizes). Dashed red lines indicate the 54 effective iterations suggested by GEMS. Each row has different maximum y-axis scaling to better illustrate the differences between results for different sphere sizes.

Figure 9.6 demonstrates voxel-to-voxel noise is largely independent of ROI size when no filtering is applied (top row) but is dependent on ROI size when Gaussian filtering is employed (bottom row). This is not an unexpected result for this type of noise measurement (Image Roughness is the average of the 60 ROI COVs): if a combination of filter width and ROI size is chosen such that the filter width is very large with respect to the ROI, one would expect little or no deviation amongst the filtered voxels inside the ROI. Voxel-to-voxel noise within each ROI would therefore be null. The use of small ROIs to measure voxel-to-voxel noise are therefore likely to underestimate the true voxel-to-voxel noise characteristics of

the image, particularly when heavy filtering has been applied, rendering the use of ROIs size-matched to each of the spheres inappropriate. Consequently, the remainder of this chapter will only use the large annular background ROIs to calculate voxel-to-voxel noise.

Figure 9.7 plots Background Variation against effective iterations for all six ROI sizes. As with the Image Roughness results, two sets of results are shown: the top row used no Gaussian filter, while the bottom row used a 10mm Gaussian filter.

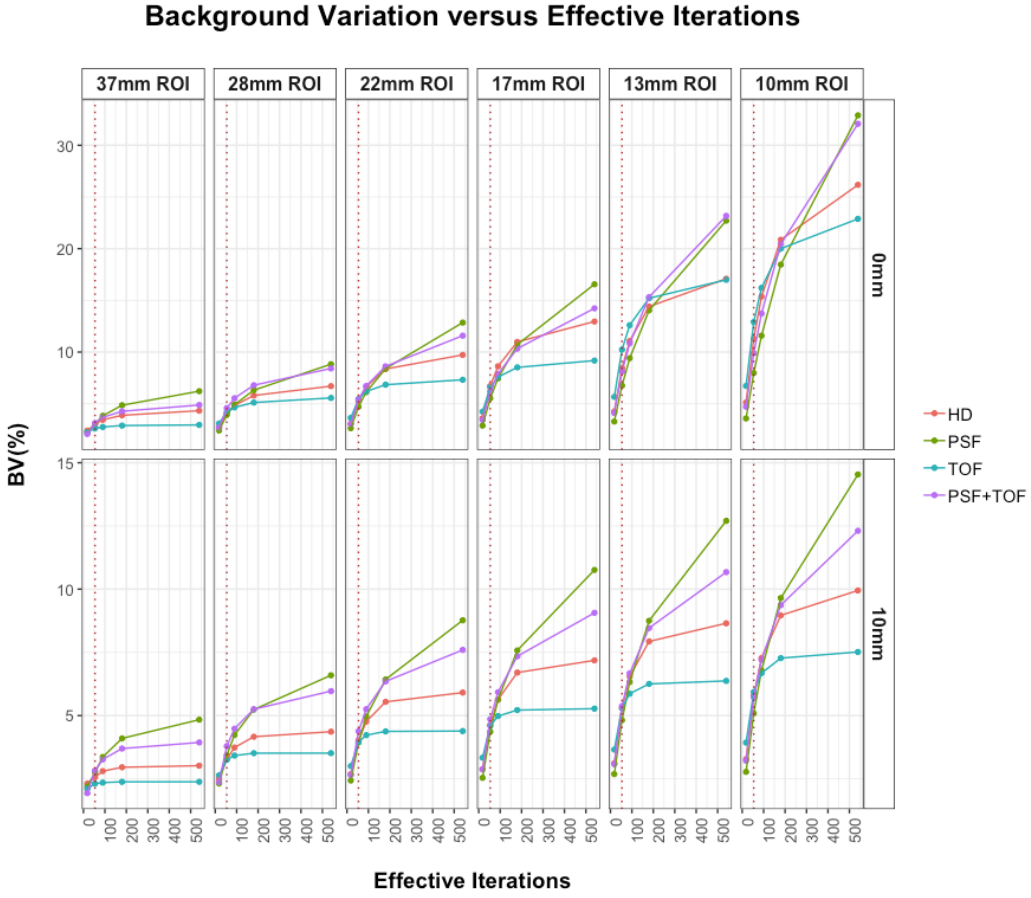


Figure 9.7: Background Variation (BV) versus effective iterations

Top row has no Gaussian filter applied, bottom row has 10mm Gaussian filter applied. Results are split by ROI size (which are matched to sphere sizes). Dashed red lines indicate the 54 effective iterations suggested by GEMS. Each row has different maximum y-axis scaling to better illustrate the differences between results for different sphere sizes.

Background Variation results are in contrast to the Image Roughness results. Background Variation is affected by ROI size for all filter widths; as ROI size decreases, Background Variation increases. This is as expected for this type of noise measurement, which is calculated by taking the COV of all 60 ROI means. As the ROI size is reduced, there is likely to be more variation among the means of different ROIs throughout the image, causing region-to-region noise to increase. The use of larger ROIs may therefore underestimate the true effects of region-to-region noise upon visualisation of smaller lesions, regardless of

image filtering. The use of ROIs size-matched to each of the spheres is therefore appropriate for region-to-region noise measurements.

The difference in behaviour between Image Roughness and Background Variation with respect to ROI size raises the following question: which noise metric is more appropriate to consider when assessing lesion detection? The literature shows that studies performing SNR analysis favour voxel-to-voxel noise measurements (using large ROIs) instead of region-to-region measurements; the standard deviation of large background ROIs are typically used, regardless of sphere/lesion size [18], [111], [134], [158], [223]. However, this may not fully illustrate the effects of background noise upon lesion detection, particularly when PSF corrections are applied during reconstruction.

9.4.2.2 SNR Results

This section will therefore assess SNR_{mean} and SNR_{max} using both voxel-to-voxel noise (standard deviation of annular background regions) and region-to-region noise (Background Variation calculated using ROIs matched to sphere size); to the author's knowledge, the use of Background Variation in SNR calculations is a novel approach. This section will focus on the smallest sphere results; however, largest sphere results will be included for comparison. Correlation between SNR results and the human observer study will also be assessed.

Figure 9.8 plots voxel-to-voxel based SNR_{mean} against effective iterations for the largest (top row) and smallest (bottom row) spheres. Each column in Figure 9.8 represents a different Gaussian filter width.

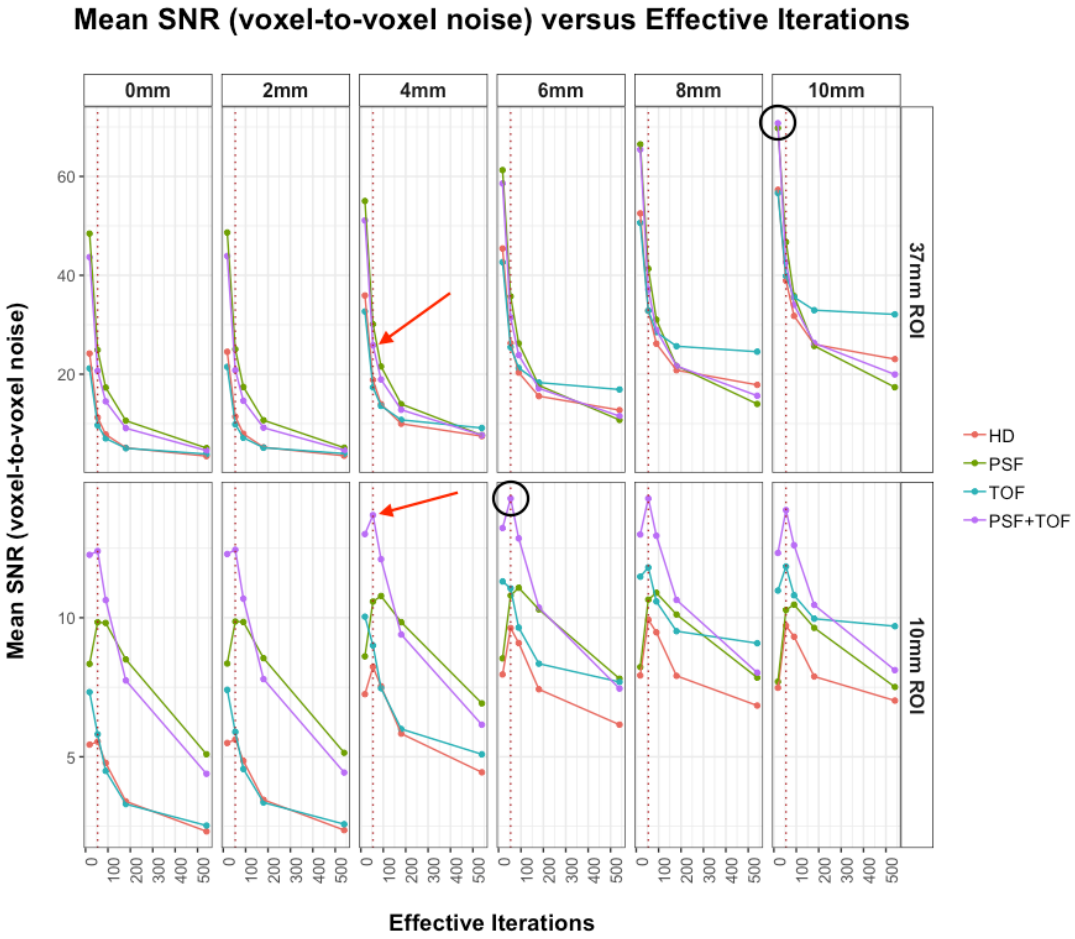


Figure 9.8: SNR_{mean} results (using voxel-to-voxel noise measurements) for largest and smallest spheres, versus effective iterations

Each column represents a different filter width, increasing from left to right. Dashed red line denotes 54 effective iterations and red arrows denote PSF+TOF and 4mm filter, suggested by GEMS for clinical reconstructions. Black circles indicate maximised SNR results.

Figure 9.8 demonstrates the smallest sphere's SNR_{mean}, when calculated using a voxel-to-voxel noise metric, is maximised by using PSF+TOF with 54 effective iterations and 6mm filter (SNR_{mean} = 14.3), as indicated by the black circle. In contrast, the largest sphere SNR_{mean} is maximised when PSF+TOF is applied with 18 effective iterations and a 10mm filter (SNR_{mean} = 70.8).

Figure 9.9 plots voxel-to-voxel based SNR_{max} against effective iterations for the largest and smallest spheres.

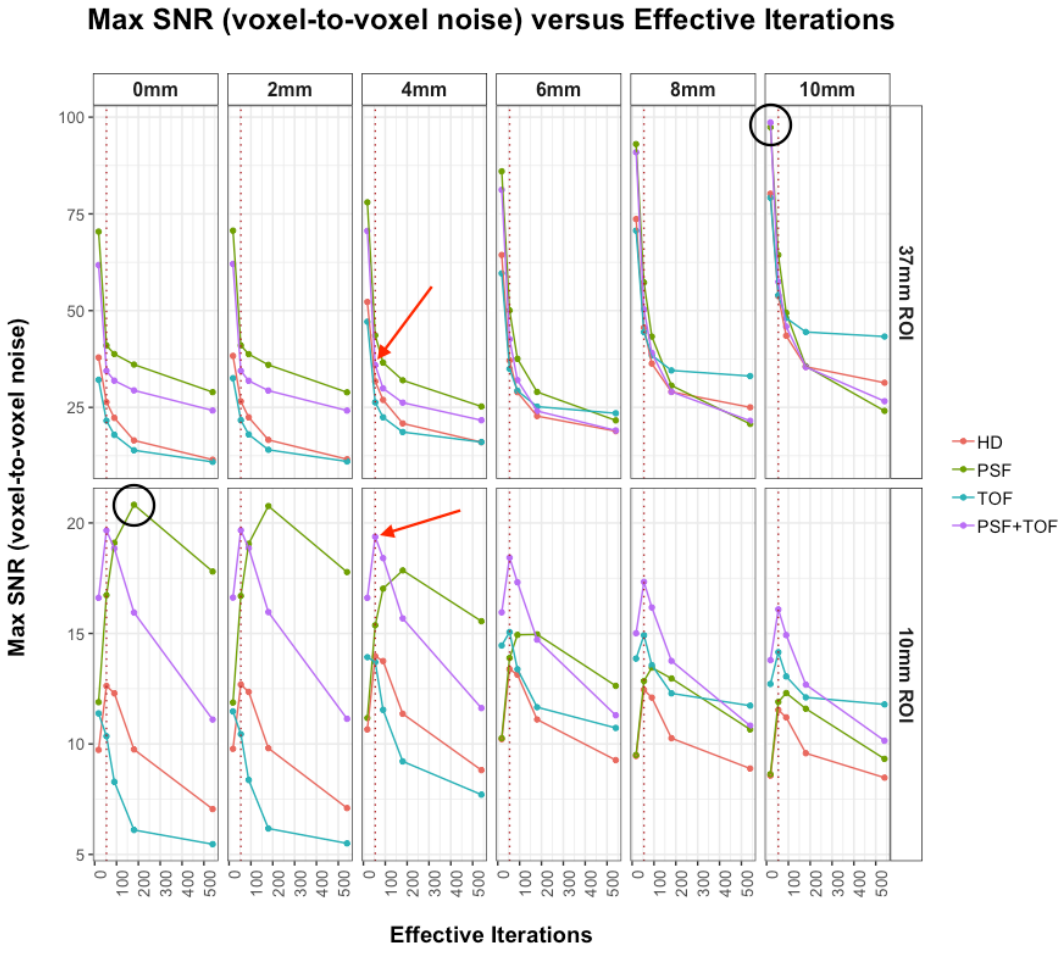


Figure 9.9: SNR_{max} results (using voxel-to-voxel noise measurements) for largest and smallest spheres, versus effective iterations

Each column represents a different filter width, increasing from left to right. Dashed red line denotes 54 effective iterations and red arrows denote PSF+TOF and 4mm filter, suggested by GEMS for clinical reconstructions. Black circles indicate maximised SNR results.

Figure 9.9 demonstrates the smallest sphere's SNR_{max} is maximised by using PSF-only with 180 effective iterations and no Gaussian filter (SNR_{max} = 20.8). This is in contrast with the SNR_{mean} results, which required a different combination of reconstruction parameters for the maximum result. SNR_{max} of the largest sphere is maximised when PSF+TOF is applied with 18 effective iterations and a 10mm filter (SNR_{max} = 98.6). The same combination of reconstruction parameters therefore maximises both SNR_{mean} and SNR_{max} of the largest sphere.

Figure 9.10 plots region-to-region based SNR_{mean} against effective iterations for the largest (top row) and smallest (bottom row) spheres.

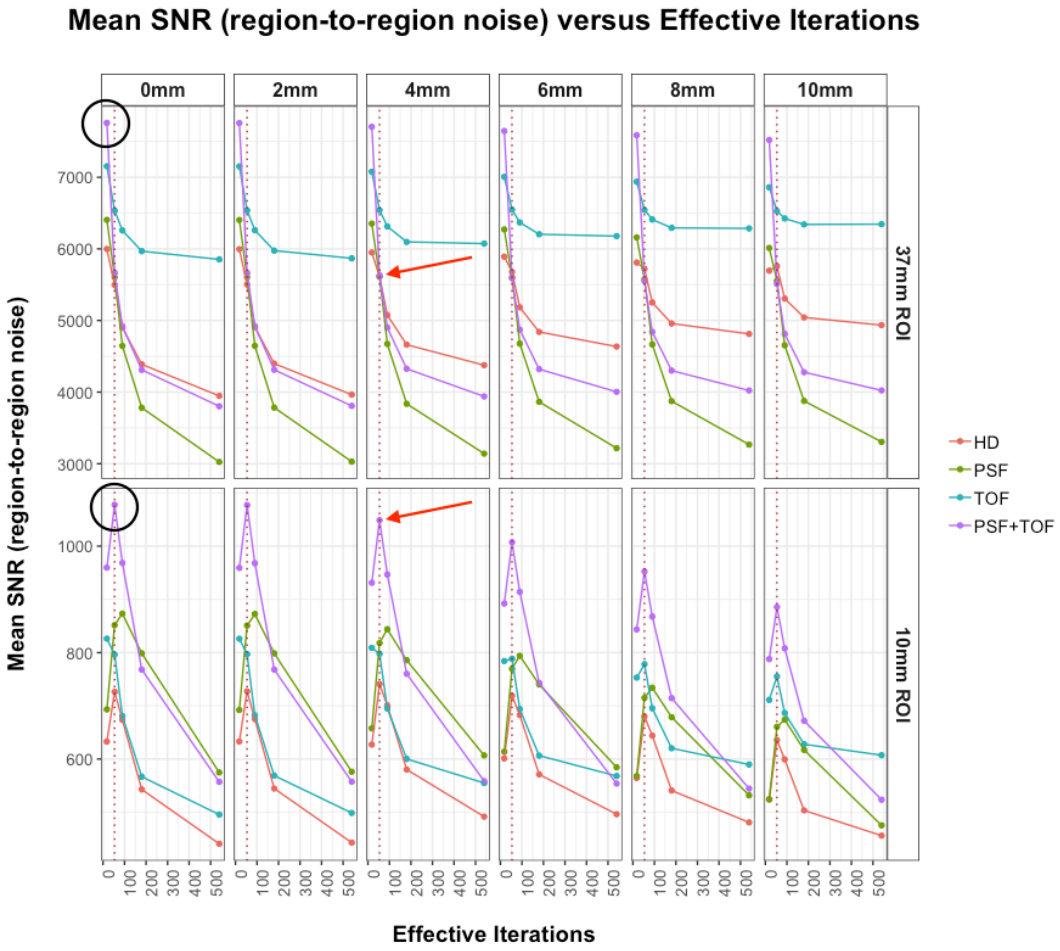


Figure 9.10: SNR_{mean} results (using region-to-region noise measurements) for largest and smallest spheres, versus effective iterations

Each column represents a different filter width, increasing from left to right. Dashed red line denotes 54 effective iterations and red arrows denote PSF+TOF and 4mm filter, suggested by GEMS for clinical reconstructions. Black circles indicate maximised SNR results.

Figure 9.10 demonstrates the smallest sphere's SNR_{mean}, when calculated using a region-to-region noise metric, is maximised by using PSF+TOF with 54 effective iterations and no Gaussian filter (SNR_{mean} = 1,078). The largest sphere SNR_{mean} is maximised when PSF+TOF is applied with 18 effective iterations and no Gaussian filter (SNR_{mean} = 7,759).

Finally, Figure 9.11 plots region-to-region based SNR_{max} against effective iterations for the largest (top row) and smallest (bottom row) spheres.

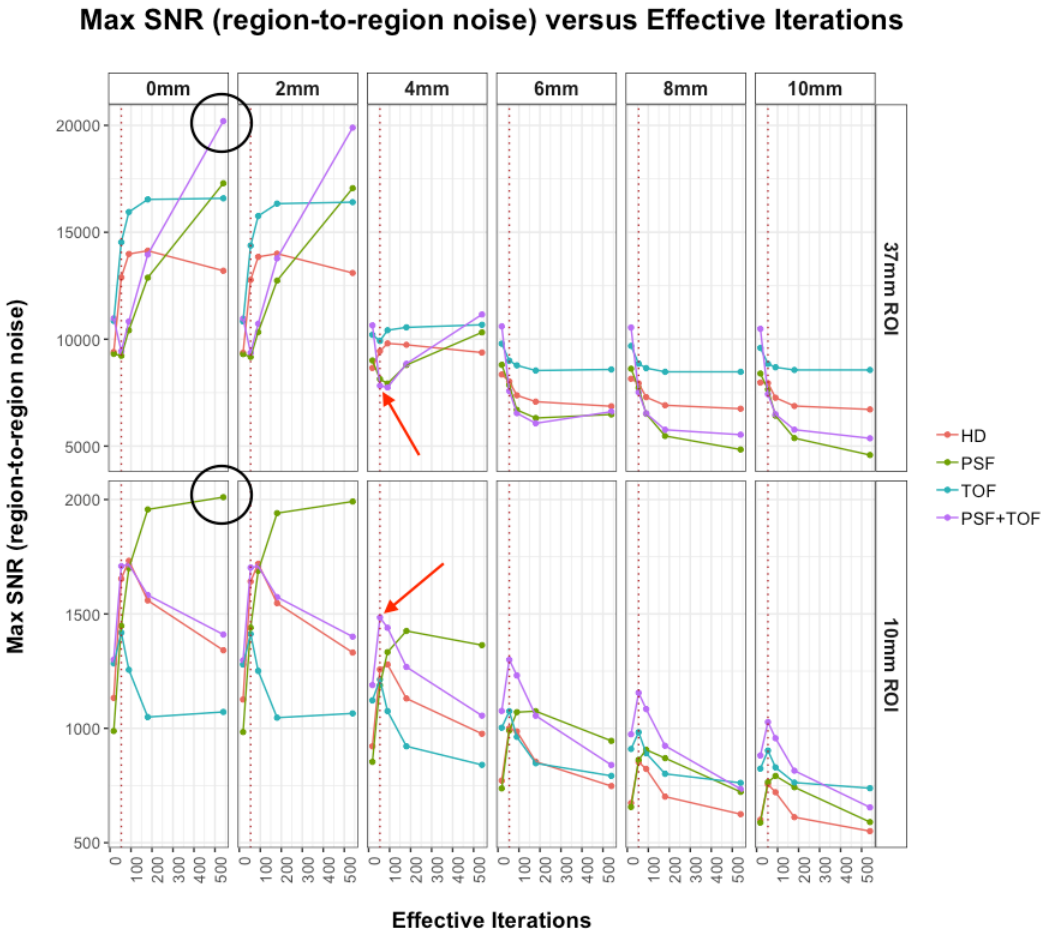


Figure 9.11: SNR_{max} results (using region-to-region noise measurements) for largest and smallest spheres, versus effective iterations

Each column represents a different filter width, increasing from left to right. Dashed red line denotes 54 effective iterations and red arrows denote PSF+TOF and 4mm filter, suggested by GEMS for clinical reconstructions. Black circles indicate maximised SNR results.

Figure 9.11 demonstrates the smallest sphere's SNR_{max} is maximised when using PSF-only with 540 effective iterations and no Gaussian filter (SNR_{max} = 2,010). As with the voxel-to-voxel SNR results, the SNR_{mean} and SNR_{max} results require different combinations of reconstruction parameters to achieve the maximum result. SNR_{max} of the largest sphere is maximised when PSF+TOF is applied with 540 effective iterations and no Gaussian filter (SNR_{max} = 20,201). The graph suggests the SNR_{max} would increase further if more effective iterations were applied. In contrast to the voxel-to-voxel SNR results, SNR_{mean} and SNR_{max} results based on region-to-region noise require different combinations of reconstruction parameters to achieve the maximum result.

Figure 9.12 compares phantom images of the four reconstructions found to maximise the four different SNR measurements for the smallest sphere.

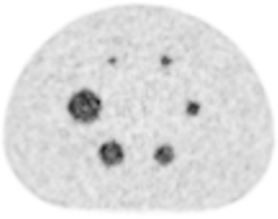
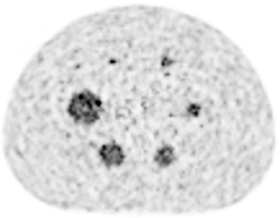
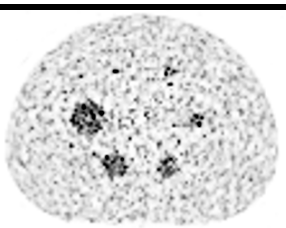
	Voxel-to-Voxel Noise	Region-to-Region Noise
SNR_{mean}		
	SNR_{mean} (voxel-to-voxel noise): PSF+TOF, 6mm FWHM Filter, 54 Effective Iterations	SNR_{mean} (region-to-region noise): PSF+TOF, no Gaussian Filter, 54 Effective Iterations
SNR_{max}		
	SNR_{max} (voxel-to-voxel noise): PSF-only, no Gaussian Filter, 180 Effective Iterations	SNR_{max} (region-to-region noise): PSF-only, no Gaussian Filter, 540 Effective Iterations

Figure 9.12: Phantom images with maximised small sphere SNR

Both SNR_{mean} results for the smallest sphere were maximised when PSF+TOF was applied with 54 effective iterations. These results are reasonably close to the parameters suggested by GEMS for clinical imaging (same reconstruction method and effective iterations, with a 4mm filter instead of 0mm/6mm shown here). However, these results are not consistent with the qualitative results from the observer study (Section 9.4.1).

Both SNR_{max} results appear to be dominated by the effect of PSF upon the maximum voxel in the smallest sphere, which has allowed a greater level of background noise to be present. One can see that some voxels in the background of these phantom images appear to have similar activity concentrations to voxels within the spheres; this is most evident for 540 effective iterations. Noise within the image therefore increases the risk of false positive lesions being identified within the uniform background.

The phantom images in Figure 9.12 can be compared with each other and with those in Figure 9.13, which shows the subset of reconstructions judged to maximise smallest sphere detectability by the qualitative analysis. The results for 54 effective iterations are also included in Figure 9.13 to aid comparison with the SNR results. Only one reconstruction combination appears in both Figure 9.12 and Figure 9.13: PSF+TOF applied with 54 effective iterations and no Gaussian filter maximised SNR_{mean} when region-to-region noise was considered. However, of the nine images shown in Figure 9.13, this reconstruction does not maximise detection of the smallest sphere from a qualitative viewpoint. This

suggests that SNR results do not necessarily correlate with human observer assessments of small lesion detection.

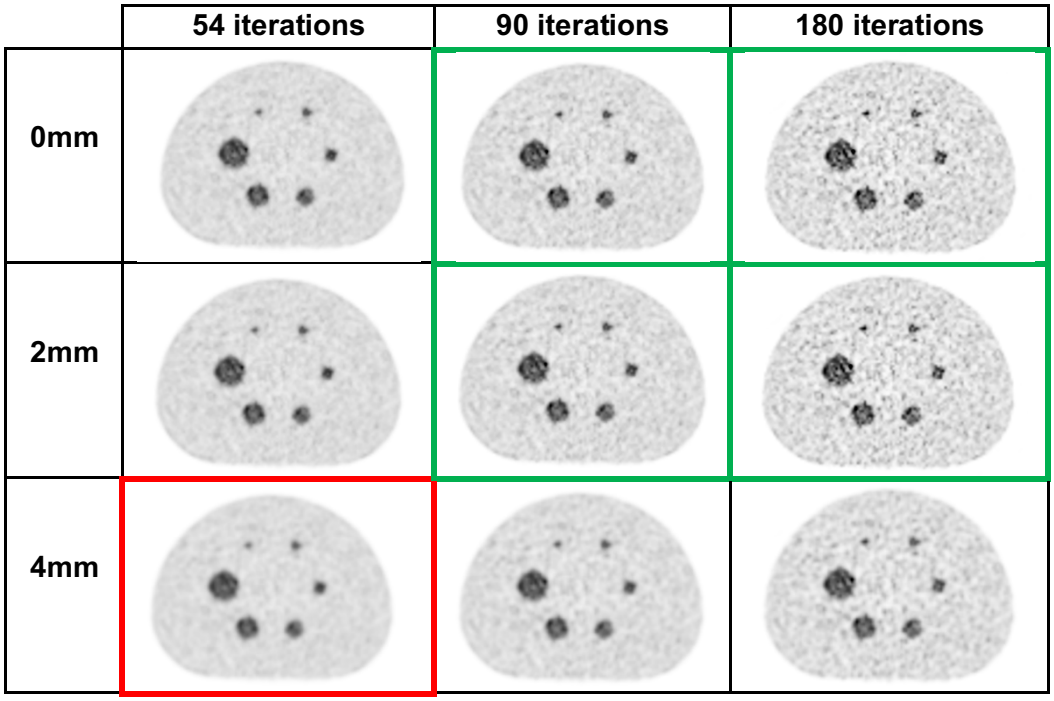


Figure 9.13: PSF+TOF phantom images judged to maximise small sphere detections

Red box denotes 54 effective iterations and 4mm Gaussian filter, suggested by GEMS for clinical reconstruction. Green boxes denote reconstructions favoured by observers.

The relationship between HCRC and noise was then assessed to determine if it had a better correlation with the human observer preferences than the SNR results.

9.4.2.3 HCRC Results

This section plots $HCRC_{mean}$ and $HCRC_{max}$ against both voxel-to-voxel and region-to-region noise. This type of dual-metric trade-off curve analysis is widely used in the literature when comparing different reconstruction techniques [15], [16], [25], [26], [49], [90], [227], [92], [130], [133], [169], [189], [224]–[226], [43], [37]. As with the SNR section, this section focuses on the smallest sphere results; however, largest sphere results are included for comparison. Curves in these plots demonstrate the increase in both HCRC and noise for each effective iteration increment. In each plot, one can attempt to select the data point that demonstrates the optimal balance of HCRC on the y-axis (as large as possible) and noise on the x-axis (as low as possible). Ideally, one would select a combination that produces a data point as close as possible to the top-left corner of the plot (maximising HCRC while minimising noise).

Figure 9.14 plots $HCRC_{mean}$ against voxel-to-voxel noise for the largest (top row) and smallest (bottom row) spheres. Each point in the graph represents a different number of applied effective iterations (corresponding to 18, 54, 90, 180 and 540 effective iterations when viewed left to right).

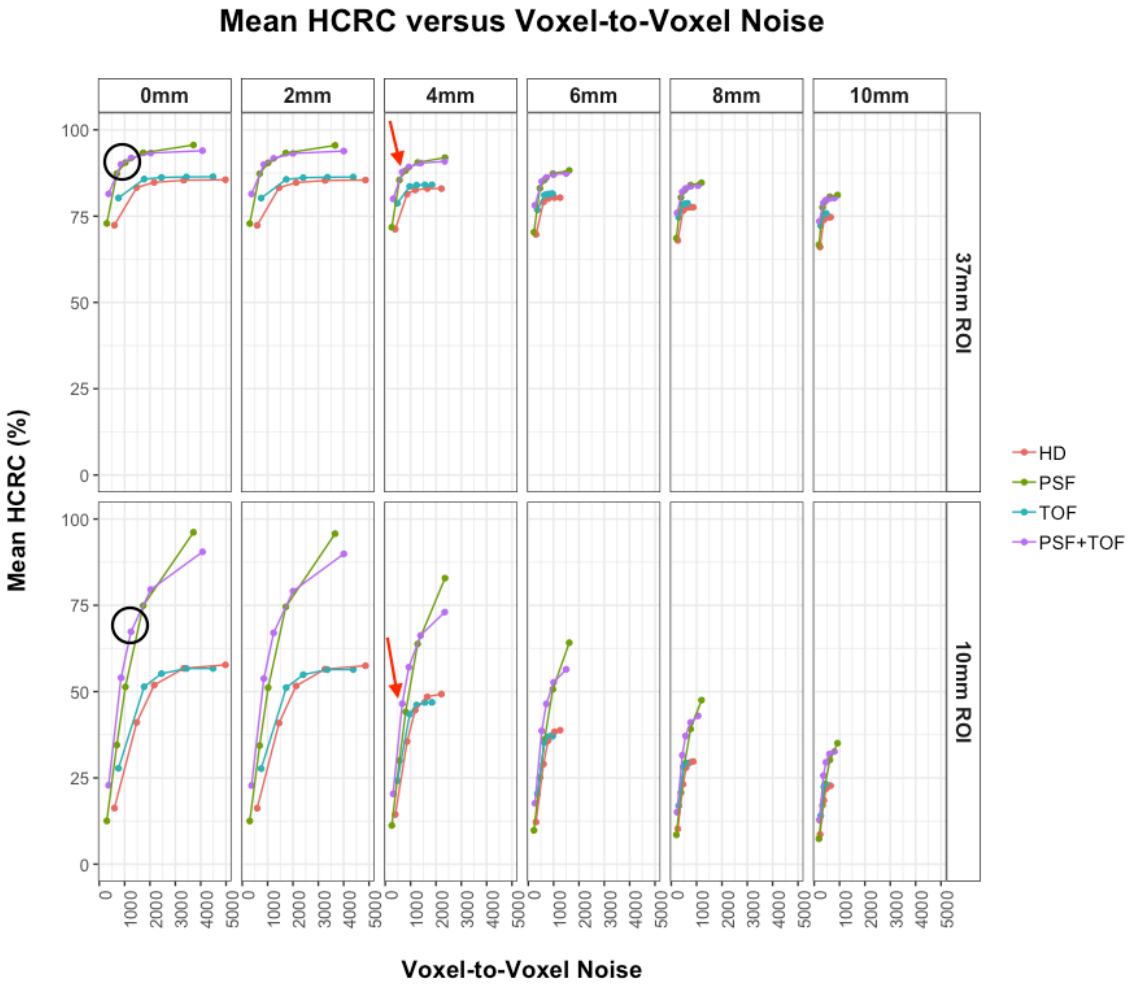


Figure 9.14: $HCRC_{mean}$ versus voxel-to-voxel noise for largest and smallest spheres

Each column represents a different filter width, increasing from left to right. Points in graph correspond to 18, 54, 90, 180 and 540 effective iterations when viewed left to right. Red arrows denote PSF+TOF with 54 effective iterations and 4mm filter, suggested by GEMS for clinical reconstructions. Black circles indicate optimal results.

Figure 9.14 suggests the optimal reconstruction for the smallest sphere appears to be PSF+TOF when approximately 90 effective iterations are applied without any filtering (black circle on graph): $HCRC_{mean}$ = 67.3%; standard deviation in the background ROIs = 1,250.3 (this level of voxel-to-voxel noise appears to be acceptable, based on the qualitative results). Non-PSF reconstructions converge to lower $HCRC_{mean}$ results with greater voxel-to-voxel noise levels. PSF+TOF generally produces lower voxel-to-voxel noise results than PSF-only, with greater corresponding $HCRC_{mean}$ results. Large sphere $HCRC_{mean}$ also

appears to be maximised by PSF+TOF when approximately 90 effective iterations are applied without any filtering.

Figure 9.15 plots $HCRC_{max}$ against voxel-to-voxel noise for the largest and smallest spheres. It should be noted that while the ideal HCRC result is 100%, $HCRC_{max}$ often exceeds this, particularly when image noise is high and partial volume effects are limited (i.e. the sphere/lesion is large).

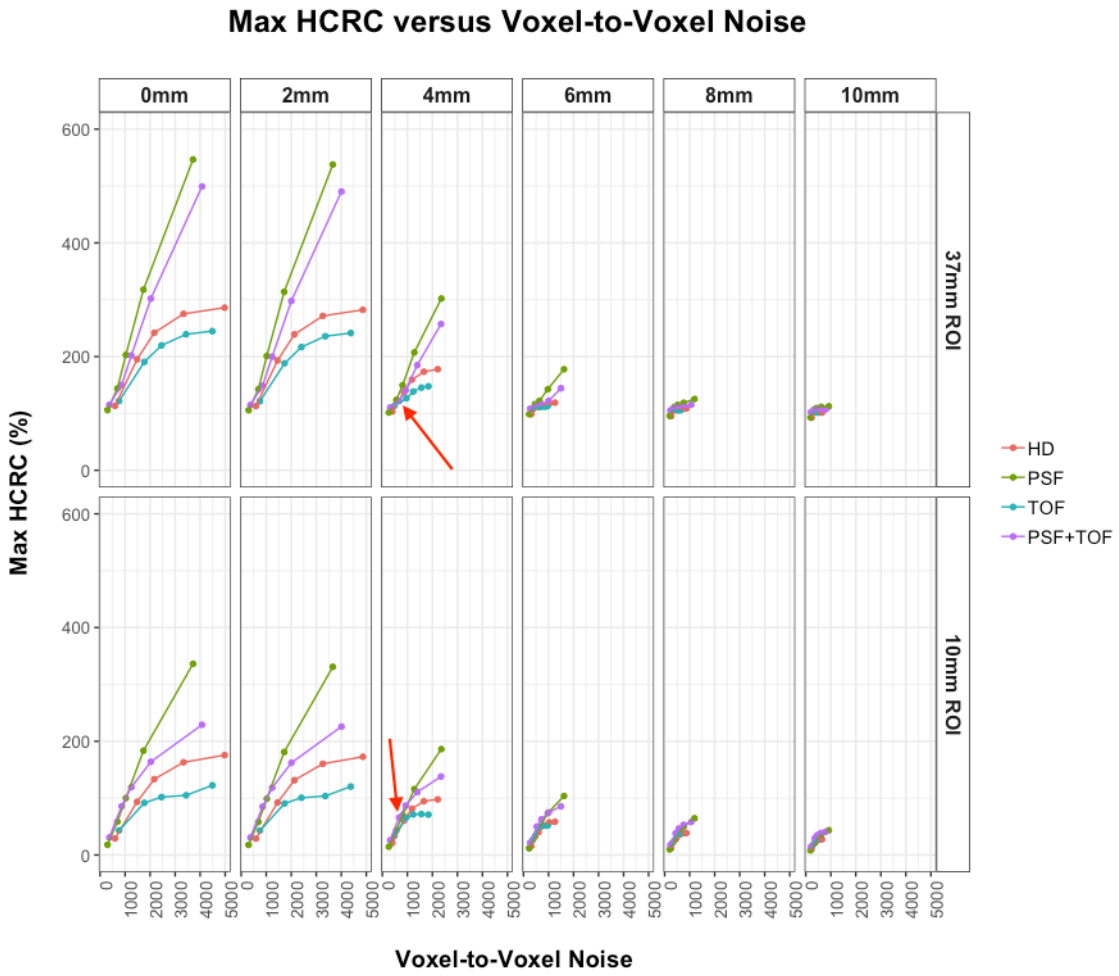


Figure 9.15: $HCRC_{max}$ versus voxel-to-voxel noise for largest and smallest spheres
Each column represents a different filter width, increasing from left to right. Points in graph correspond to 18, 54, 90, 180 and 540 effective iterations when viewed left to right. Red arrows denote PSF+TOF with 54 effective iterations and 4mm filter, suggested by GEMS for clinical reconstructions.

There appears to be almost linear relationships between voxel-to-voxel noise and $HCRC_{max}$, particularly at lower iterations; this is not surprising, as results based on a single, maximum voxel will be inherently dependent upon image noise, as discussed previously in this thesis. It is difficult to use $HCRC_{max}$ to suggest the optimum reconstruction parameter combination.

Figure 9.16 plots $HCRC_{mean}$ against region-to-region noise for the largest and smallest spheres.

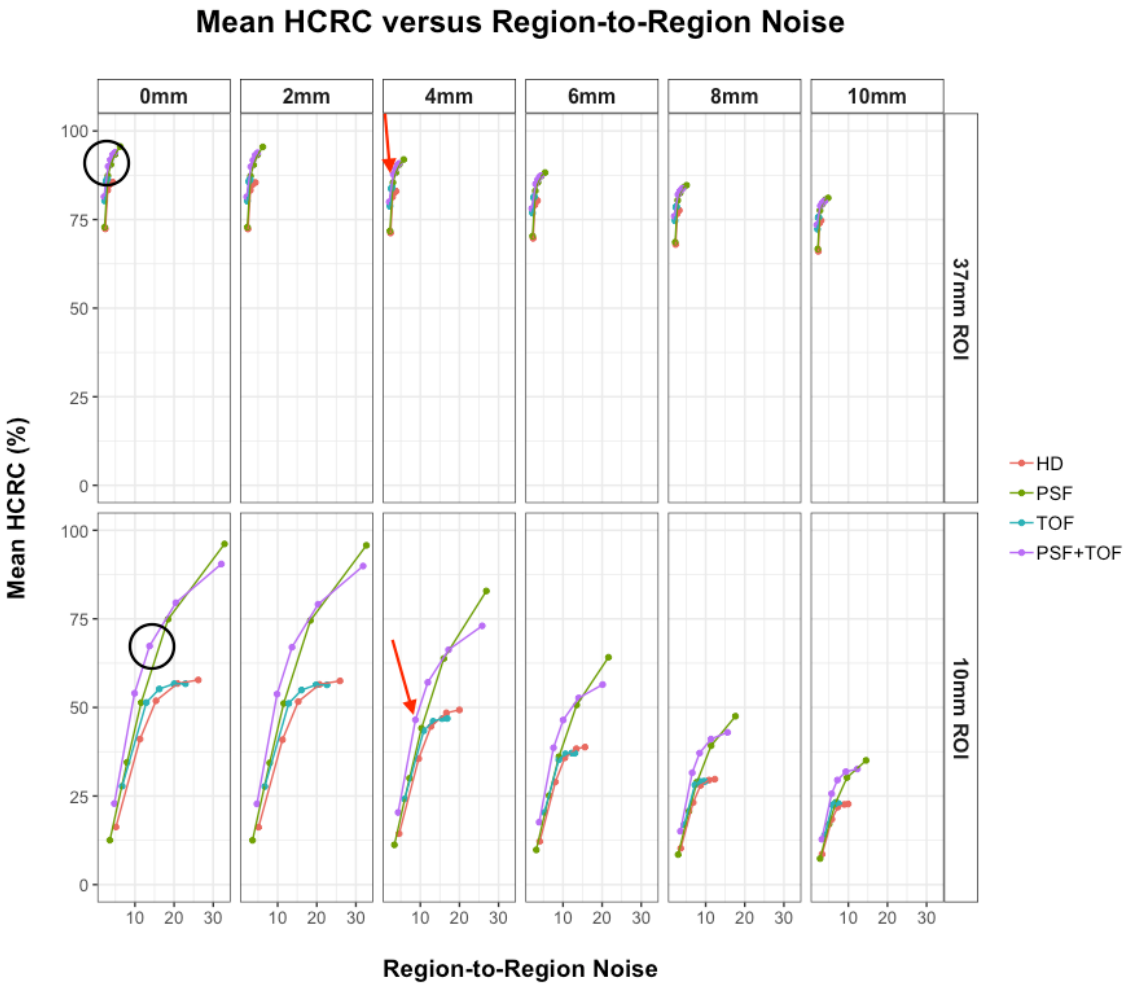


Figure 9.16: $HCRC_{mean}$ versus region-to-region noise for largest and smallest spheres

Each column represents a different filter width, increasing from left to right. Points in graph correspond to 18, 54, 90, 180 and 540 effective iterations when viewed left to right. Red arrow denotes PSF+TOF with 54 effective iterations and 4mm filter, suggested by GEMS for clinical reconstructions. Black circles indicate optimal results.

As with the voxel-to-voxel noise results, Figure 9.16 suggests the optimal reconstruction for the smallest sphere appears to be PSF+TOF when approximately 90 effective iterations are applied without any filtering (black circle): $HCRC_{mean}$ = 67.3%; BV = 13.7% (again, this level of region-to-region noise appears to be acceptable, based on the qualitative results). Non-PSF reconstructions converge to lower $HCRC_{mean}$ results, with generally lower region-to-region noise levels. PSF+TOF generally produces lower region-to-region noise results than PSF-only, with greater corresponding $HCRC_{mean}$ results. As with the voxel-to-voxel noise results, large sphere $HCRC_{mean}$ also appears to be maximised by PSF+TOF when approximately 90 effective iterations are applied without any filtering.

Finally, Figure 9.17 plots $HCRC_{max}$ against region-to-region noise for the largest and smallest spheres.

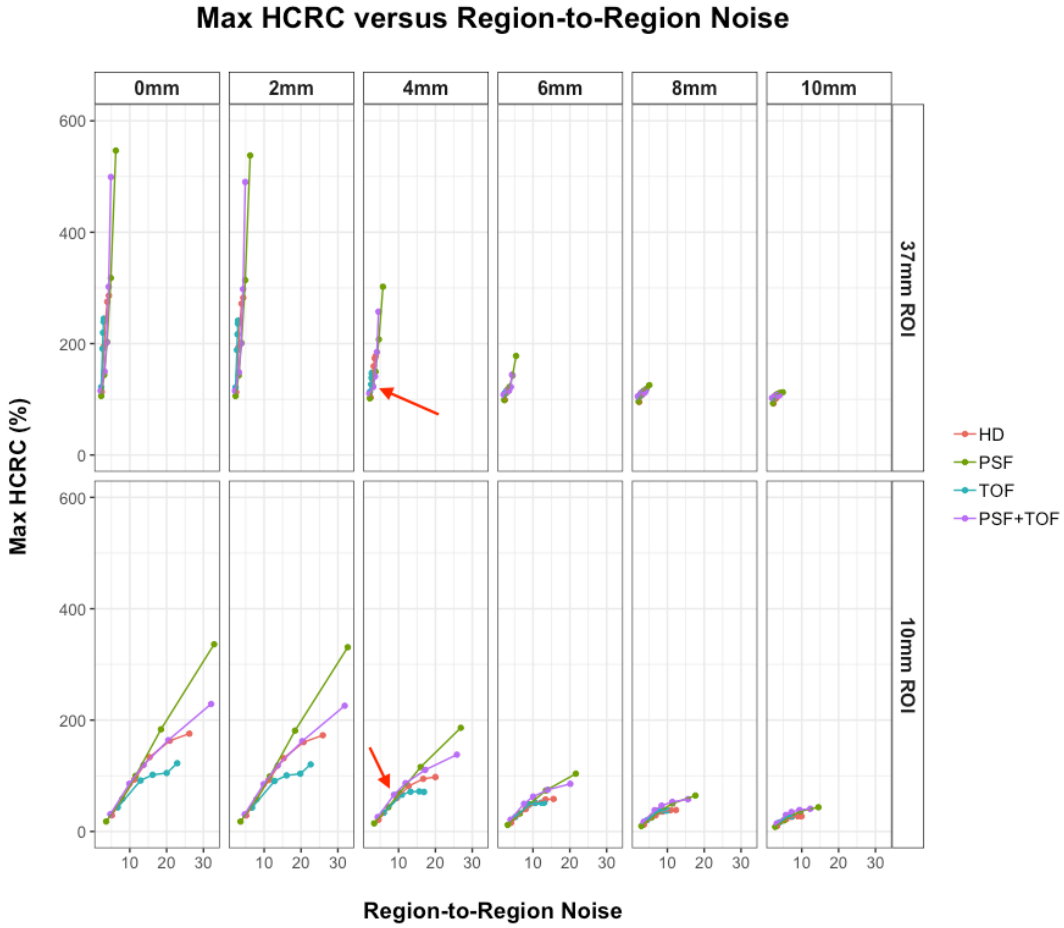


Figure 9.17: $HCRC_{max}$ versus region-to-region noise for largest and smallest spheres

Each column represents a different filter width, increasing from left to right. Points in graph correspond to 18, 54, 90, 180 and 540 effective iterations when viewed left to right. Red arrow denotes PSF+TOF with 54 effective iterations and 4mm filter, suggested by GEMS for clinical reconstructions.

As with the voxel-to-voxel noise results, relationships between region-to-region noise and $HCRC_{max}$ appear to be approximately linear. It is difficult to use $HCRC_{max}$ to suggest the optimum reconstruction parameter combination.

Figure 9.18 compares images of the PSF+TOF reconstructions likely to provide the optimum trade-off between $HCRC_{mean}$ and image noise (both voxel to voxel and region-to-region) for the smallest sphere. The images in Figure 9.18 are in close agreement with the qualitative observer study (preferences indicated by the green boxes).

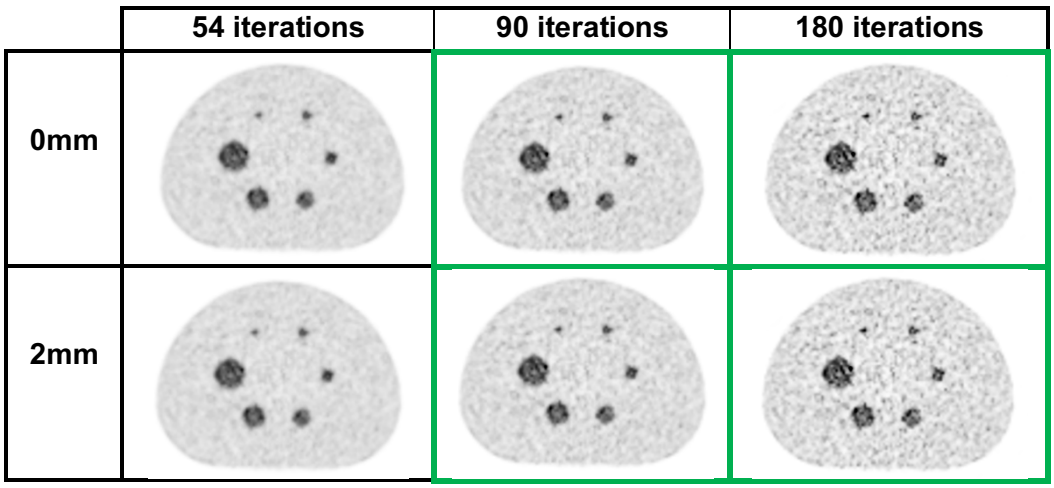


Figure 9.18: Phantom images that appeared to produce optimum small sphere $HCRC_{mean}$ and noise trade-off

Green boxes denote reconstructions favoured by observers.

In summary, $HCRC_{mean}$ versus noise analysis appears to provide a better prediction of the human observer preferences than SNR analysis when detection of the smallest sphere is the primary consideration. Part 1 of this chapter therefore concludes the following reconstruction parameters should be applied in order to maximise detection of the smallest sphere within the body phantom:

- PSF+TOF reconstruction
- Little or no filtering
- Approximately 90 effective iterations, but potentially as many as 180 effective iterations

9.5: Part 2 Results: Narrowing the Range of Clinically Relevant Parameter Combinations

The qualitative human observer preferences and the quantitative $HCRC_{mean}$ versus noise analysis in the previous section clearly demonstrated certain choices of parameters are unsuitable for use when attempting to maximise small sphere detection. However, the incremental increases in both the Gaussian filter width and, more particularly, the number of effective iterations, were relatively coarse; the optimal combination of reconstruction parameters for small sphere detection may not yet have been examined. A second group of reconstructions were therefore assessed briefly, as detailed in Table 9.3. Only PSF+TOF is used, as this was shown in Part 1 to demonstrate superior performance. Effective iterations are varied from 54 up to 180 in increments of 18: i.e. 18 OSEM subsets were used with between 3 and 10 iterations. Finer increments (0.5mm) in the Gaussian filter are also used, with a maximum of 4mm. No finer increments are used between 0mm and 2mm: Chapter 6 demonstrated a filter width of at least approximately 2mm is required to affect

voxel values when 2.73mm voxel were used. Images are assessed using $HCRC_{mean}$ versus noise analysis for the smallest sphere only.

Reconstruction Parameter	Settings Used	Number of Result Groups
Reconstruction Method	PSF+TOF	1
Effective OSEM Iterations	54, 72, 90, 108, 126, 144, 162, 180	8
Gaussian Filter (FWHM)	0mm, then 2 → 4mm in 0.5mm increments	6
Z-Axis Filter	None	1
Matrix Size	256*256	1

Table 9.3: Clinically relevant parameter combinations used for body phantom reconstructions

Table 9.19 shows the central slice of all reconstructions produced in this section.

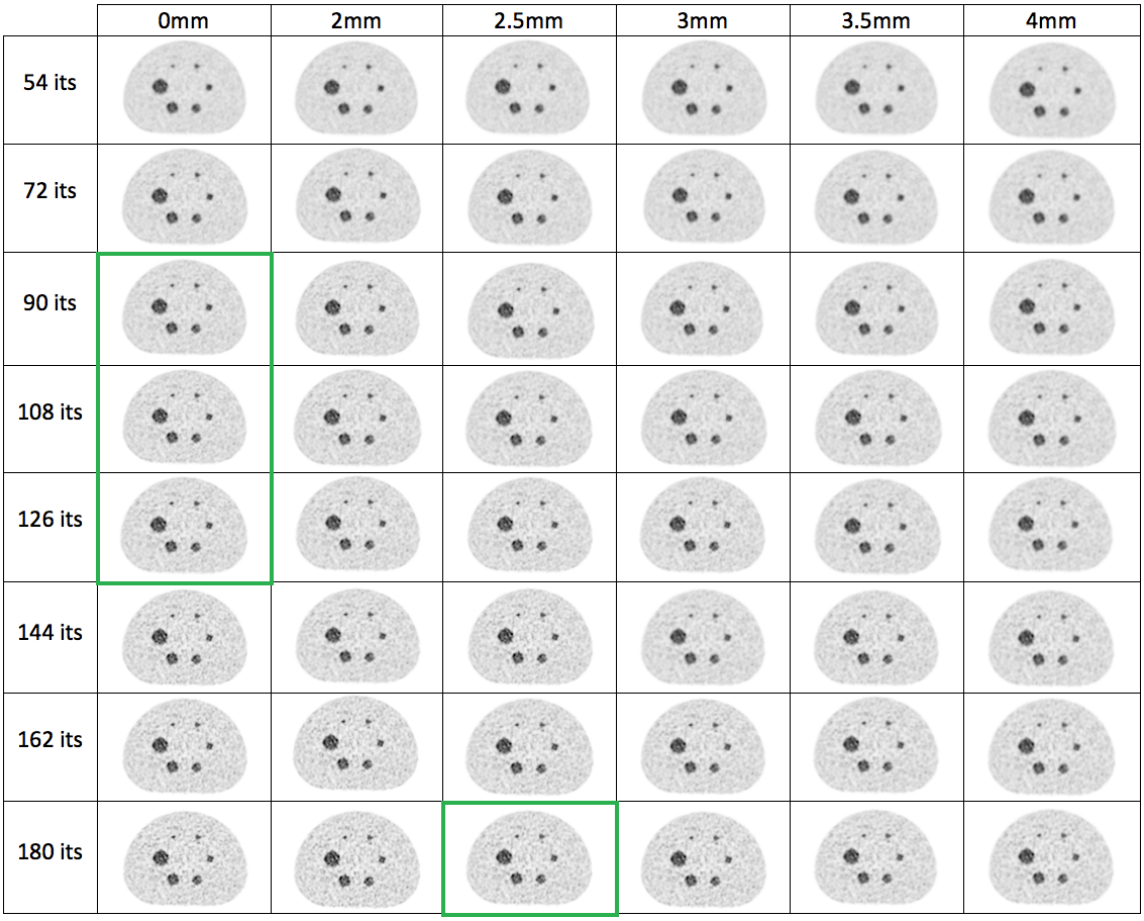


Figure 9.19: PSF+ TOF images with narrowed range of Gaussian filter widths and effective iterations

Green boxes indicate observer preferences.

Three of the four observers preferred images produced using between 90 and 126 effective iterations with no Gaussian filtering. The fourth observer preferred the images produced when 180 effective iterations were used with a 2.5mm filter.

Figure 9.20 plots $HCRC_{mean}$ versus noise for the smallest sphere only. Two different plots are shown: $HCRC_{mean}$ is plotted against both voxel-to-voxel and region-to-region noise. As before, each point in the graphs represents a different number of applied effective iterations (corresponding to 54, 72, 90, 108, 126, 144, 162 and 180 effective iterations when viewed left to right).

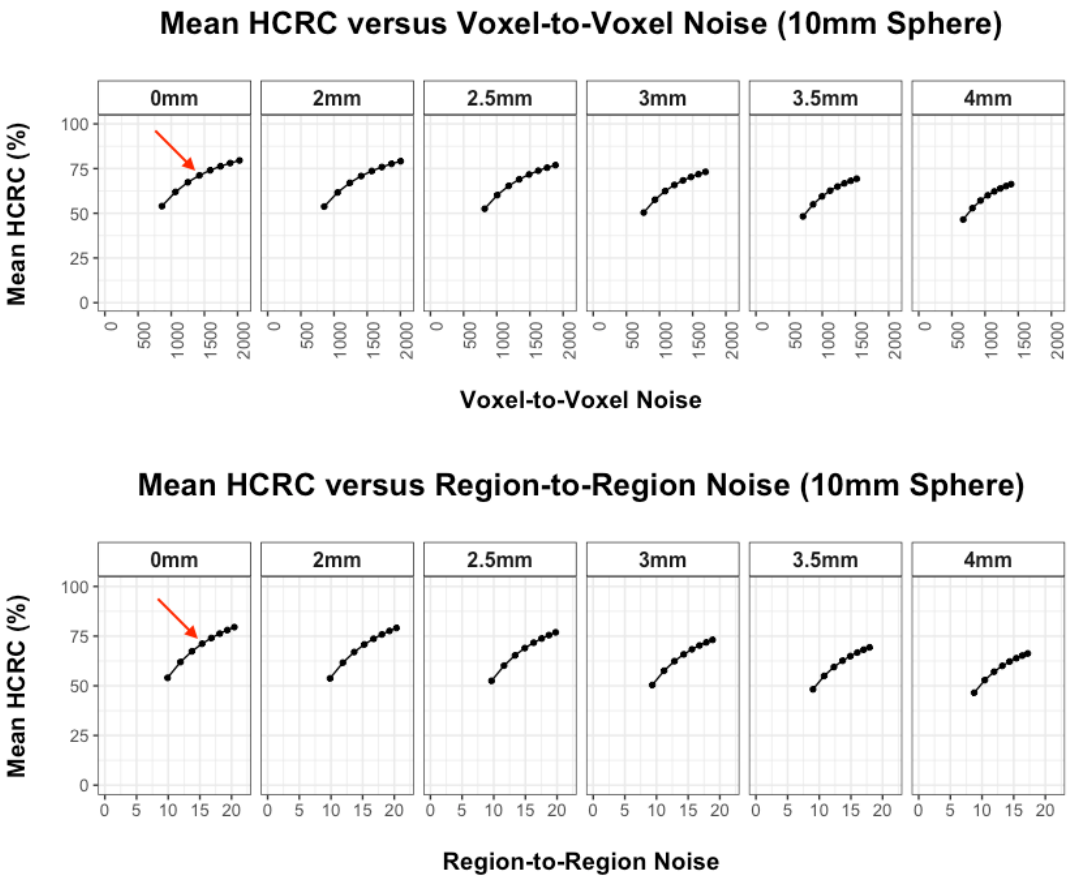


Figure 9.20: $HCRC_{mean}$ versus noise plots for smallest sphere only

Both voxel-to-voxel (top) and region-to-region (bottom) noise results are shown. Red arrows indicate likely optimal trade-off between noise and $HCRC_{mean}$.

With respect to voxel-to-voxel noise, there appears to be a diminishing $HCRC_{mean}$ return when applying more than approximately 108 effective iterations; after this, the increase in noise appears to be more significant than the increase in $HCRC_{mean}$ (particularly when little/no Gaussian filtering is applied). A similar effect is observed for the region-to-region noise results. The results in Figure 9.20 therefore suggest approximately 108 effective iterations combined with no Gaussian filter may produce the optimal $HCRC_{mean}$ results for the smallest sphere. This is in broad agreement with three of the observers, who preferred images reconstructed with between 90 and 126 effective iterations.

The use of finer effective iteration increments in this section has therefore revised the apparent optimum number of effective iterations for $HCRC_{mean}$ from 90 to 108.

9.6 Part 3 Results: Reproducibility Study

Analyses performed in Part 1 and Part 2 of this chapter were based on a single phantom acquisition; a small study was therefore undertaken to assess the reproducibility of these findings. Five separate acquisitions of the phantom, filled with similar activities and sphere-to-background ratios (as detailed in Table 9.4) were undertaken on five separate days. A single operator filled and positioned the phantom for all five acquisitions, which were then reconstructed using the same set of reconstruction parameters used in Part 2 of this chapter (detailed in Table 9.3). This ensured reproducibility was assessed over a clinically relevant range of reconstruction parameters.

	Background Activity Concentration	Sphere Activity Concentration	Ratio
Phantom 1	5,263Bq/ml	24,188Bq/ml	4.60:1
Phantom 2	5,297Bq/ml	25,100Bq/ml	4.74:1
Phantom 3	5,346Bq/ml	24,830Bq/ml	4.64:1
Phantom 4	5,290Bq/ml	21,601Bq/ml	4.08:1
Phantom 5	5,296Bq/ml	20,750Bq/ml	3.92:1

Table 9.4: Body phantom activity concentrations for reproducibility study

Background activity concentrations at the time of scanning were similar for all five phantoms: the difference between the maximum and minimum background activity concentrations was 83Bq/ml (1.6%). Noise measurements should therefore be comparable across all five phantoms, as all phantoms were acquired using four-minute bed-times. There was more variation amongst the sphere activity concentrations, and therefore in the sphere-to-background ratios: the difference between the maximum and minimum background activity concentrations was 4,350Bq/ml (18.7% of average sphere activity concentration). Only $HCRC_{mean}$ results were used to assess sphere reproducibility: these have been shown to be the most relevant quantitative result for small sphere detection and should still be comparable in spite of absolute differences in sphere activity concentrations.

Figure 9.21 shows the voxel-to-voxel noise reproducibility results. As discussed previously, this type of noise measurement is not related to sphere size. Error bars denote 95% confidence intervals.

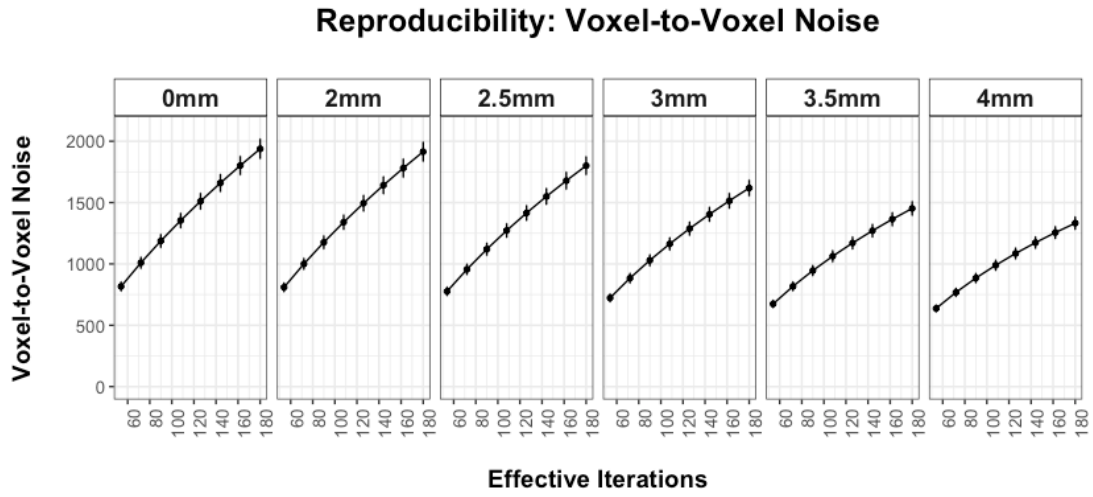


Figure 9.21: Voxel-to-voxel noise reproducibility results
Error bars show 95% confidence intervals.

Error bars are small in proportion to result magnitudes, even for the noisiest reconstruction; when 180 effective iterations are applied with no filtering, the result is 1,938 ± 82 (i.e. ± 4.2%). This demonstrates that voxel-to-voxel noise results are consistent across the five phantoms acquisitions, as expected. ROIs used to measure voxel-to-voxel noise are small in comparison to the phantom's background volume; small inconsistencies in ROI placement between the phantoms are therefore highly unlikely to influence the results.

Figure 9.22 shows the region-to-region noise reproducibility results for the largest and smallest spheres.

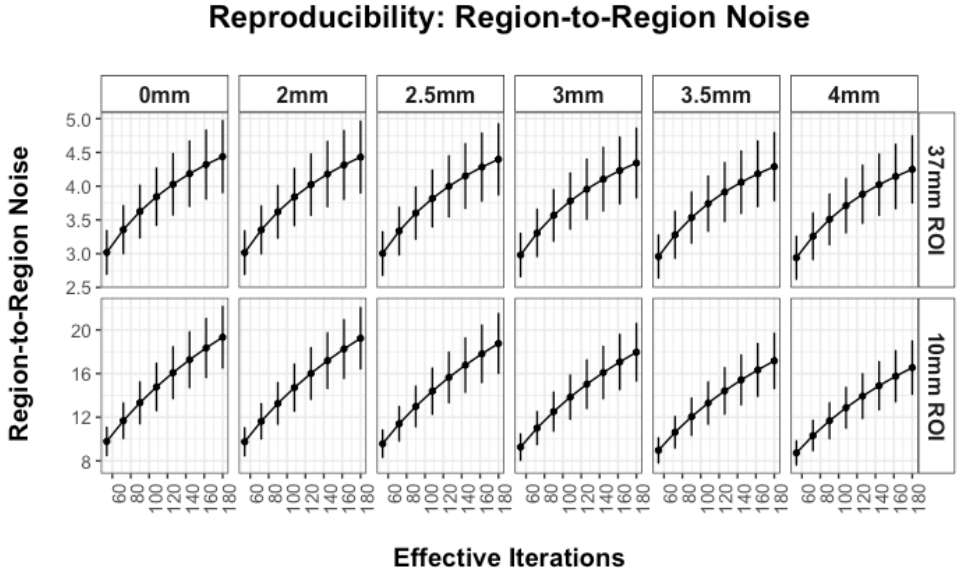


Figure 9.22: Region-to-region noise reproducibility results for largest and smallest spheres
Error bars show 95% confidence intervals.

The largest errors for both spheres are produced when 180 effective iterations are used with no filtering, as one would expect: largest sphere result is 4.44 ± 0.54 (i.e. ± 12.2%) and the smallest sphere result is 19.33 ± 2.87 (i.e. ± 14.8%). Region-to-region noise of smaller spheres are inherently greater than that of larger spheres, as demonstrated earlier in this chapter; however, confidence intervals for both spheres are similar when considered as a proportion of the measured noise for each reconstruction. Region-to-region noise measurements are also more likely to be affected by inconsistencies in ROI placement due to the smaller size of ROIs used and variations in the background caused by the reconstruction process [90], [92].

Finally, Figure 9.23 shows the $HCRC_{mean}$ results for the largest and smallest spheres.

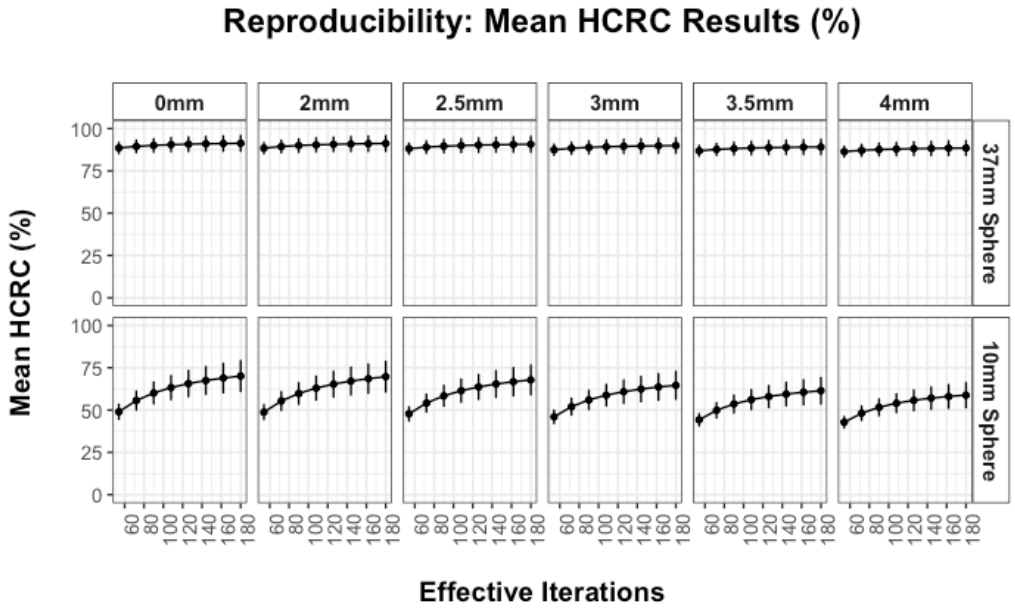


Figure 9.23: $HCRC_{mean}$ reproducibility results for largest and smallest spheres Error bars show 95% confidence intervals.

One would expect large sphere $HCRC_{mean}$ results to be reasonably consistent across the 5 phantoms as they are less sensitive to positioning inconsistencies than the small sphere (both of the phantom and of the ROIs). Figure 9.23 demonstrates this is indeed the case: for the noisiest reconstruction (180 effective iterations, no filtering), the largest sphere's result was 91.3 ± 5.0 (± 5.49% of mean value), whilst the smallest sphere's result was 70.1 ± 9.4 (± 13.4% of mean value).

All five phantom acquisition demonstrated similar noise and $HCRC_{mean}$ trends, and the observed errors are as one would expect given the nature of the measurements. The reproducibility study demonstrated that errors associated with the phantom analysis do not affect any of the conclusions reached in Parts 1 and 2 of this chapter.

9.7 Discussion

This chapter aimed to determine the optimum combination of reconstruction parameters for detecting small liver lesions. A body phantom was used to mimic lesions of varying sizes within a patient liver and the effects of varying reconstruction parameters (effective iterations, Gaussian filter width, PSF and/or TOF) were assessed qualitatively and quantitatively. Part 1 assessed a wide range of these reconstruction parameters using relatively course increments for effective iterations and Gaussian filter width, and the results were used to narrow down the range of reconstruction parameters likely to optimise small lesion detection. Part 2 assessed this narrower range of reconstruction parameters, using finer increments for effective iterations and Gaussian filter width.

Four experienced PET observers assessed qualitatively Part 1 phantom images, and all concluded the use of PSF+TOF with little or no filtering produced superior smallest sphere detectability. Observers disagreed on the optimum number of effective iterations to apply; three observers preferred 90 effective iterations while one preferred 180 effective iterations. Quantitative analysis was then performed, assessing background both voxel-to-voxel and region-to-region noise in the background, and SNR and HCRC of the spheres. Both types of noise measurement were used to calculate SNR_{mean} and SNR_{max} and plotted against both $HCRC_{mean}$ and $HCRC_{max}$ results. To the author's knowledge, the use of region-to-region noise in SNR calculations was a novel approach.

Reconstructions which were shown to maximise SNR of the smallest sphere did not agree with the phantom images preferred by the observers; however, it is of interest that SNR_{mean} calculated using region-to-region noise was closer to matching the observer preferences than traditional SNR_{mean} calculated using voxel-to-voxel noise (PSF+TOF with no filtering, but only 54 effective iterations).

HCRC results were plotted against both noise measurements and used to determine the optimum trade-off between HCRC and noise for the smallest sphere. $HCRC_{max}$ results were not useful for this purpose as they tended to demonstrate linear relationships with noise. The $HCRC_{mean}$ versus noise results, however, appeared to be optimal when PSF+TOF was combined with approximately 90 effective iterations and no filtering, for both voxel-to-voxel and region-to-region noise. This combination of reconstruction parameters was in agreement with three of the four PET observers.

Part 1 therefore concluded PSF+TOF combined with little/no filtering and approximately 90 effective iterations optimised small sphere detection, both qualitatively and quantitatively.

Furthermore, this section concluded $HCRC_{mean}$ versus noise analysis closely predicted the human observer assessment, although SNR_{mean} (when calculated using region-to-region noise) was reasonably close to the human observer results. The recommendation to use PSF+TOF is consistent with several publications discussed earlier in this chapter [8], [18], [19], [111], [149], [157], [215], [217], [220]; however, these publications all recommended fewer than 90 iteration effective iterations.

As Part 1 used relatively coarse increments for both effective iterations and Gaussian filter width, it was possible the optimum reconstruction method could have been refined further. Part 2 therefore used finer increments over the range of parameters identified by Part 1 as being most useful clinically. Only PSF+TOF was used, as Part 1 ruled out the other three methods. Effective iterations were varied between 54 and 180 in finer increments (although 54 effective iterations were considered to be too few in Part 1, it was included here as it's suggested by GEMS for clinical reconstructions). The Gaussian filter was capped at 4mm; again, this was included as it's suggested by GEMS for clinical reconstructions (despite the previous section suggesting little or no filtering should be employed).

Qualitative assessment of Part 2's images demonstrated a similar difference in opinion amongst the PET observers as Part 1. Three of the four observers preferred images produced using between 90 and 126 effective iterations with no Gaussian filtering. The fourth observer preferred the images produced when 180 effective iterations were used with a 2.5mm filter. However, the differences between these chosen images are minimal; the application of the 2.5mm filter may have compensated for the additional noise and sharpness associated with the use of 180 effective iterations.

Quantitative analysis for Part 2 focussed on small sphere $HCRC_{mean}$, as Part 1 had shown this to be most useful. Results suggested that when more than 108 effective iterations were applied, the increase in noise was greater than the increase in $HCRC_{mean}$. This in turn suggested 108 effective iterations combined with no filtering may be the optimum combination of parameters. This was in broad agreement with three of the four observer's choices.

A major limitation of these observer studies is the known sphere positions within the phantom: observers know where to look for the smallest sphere in each phantom image, which introduces a source of bias to the results. Further work should use a phantom with unpredictable sphere numbers, sizes and positions. This would also provide a measure of false positives within the background activity, particularly when noisier reconstructions are used in an attempt to improve spatial resolution. A further limitation of this study is that

phantom results may only be applicable to patients of a similar size to the phantom: this is investigated in the next chapter.

Finally, Part 3 assessed the reproducibility of noise and $HCRC_{mean}$ results over a range of clinically relevant reconstruction combinations. Whilst the error analysis demonstrated some variability in the measurements across five similar phantom acquisitions, the variations were as expected for experiments of this nature: there was nothing to suggest the quantitative results or conclusions in Parts 1 and 2 in this chapter were not valid.

9.8 Conclusions

Quantitative analysis of NEMA body phantom reconstructions suggests small sphere detection was maximised when both PSF and TOF were applied together, with 108 effective iterations and with no filtering applied. The results of an informal observer study largely agreed with this conclusion; however, one observer preferred the use of 180 effective iterations with a 2.5mm Gaussian filter. This chapter therefore concludes the following two reconstruction strategies should be applied to clinical liver data and assessed in a formal patient observer lesion detection study:

	Reconstruction Method	Effective Iterations	Gaussian filter FWHM	Z-axis Filter	Voxel Size
Recon 1	PSF+TOF	108	0mm	None	2.73mm
Recon 2	PSF+TOF	180	2.5mm	None	2.73mm

Table 9.5: Recommended reconstruction strategies for small lesion detection

This chapter further concludes that dual metric $HCRC_{mean}$ and noise analysis provided the best prediction of human observer preferences for small lesion detection. Furthermore, a novel method of calculating SNR using region-to-region noise was found to provide a better prediction of human observer preferences than traditional SNR measurements using voxel-to-voxel noise. These methods of quantitative assessments should be included when optimising PET reconstruction techniques for other clinical applications.

Chapter 10 : Comparison of NEMA Body Phantom with Patient Liver Data

The previous chapter performed multiple reconstructions of the NEMA IEC body phantom using different combinations of reconstruction parameters. Qualitative and quantitative analyses were performed with a view to identifying which combinations of parameters optimised the detection of the smallest sphere within the phantom. Two specific combinations of reconstruction parameters were identified as meriting further evaluation in a patient liver lesion detection study. However, a limitation of Chapter 9 was that results may only be applicable to patients of a similar size to the phantom. A brief study was therefore undertaken to compare the phantom to a range of patient sizes in order to predict how well the results of Chapter 9 would apply to patient liver data.

10.1 Introduction

The NEMA IEC body phantom, used throughout this thesis and described in detail in Chapter 3, is recommended for use in the evaluation of reconstructed image quality in whole-body PET imaging [228]; for example, when acceptance testing new PET imaging systems [70]. It has also been widely used by publications assessing various aspects of PET image acquisition and reconstruction, many of which have been referenced throughout this thesis. However, a single phantom can only represent a limited range of imaging situations. Patients undergoing ^{18}F-FDG imaging come in a range of weights and sizes, and therefore cannot be fully represented by the NEMA phantom.

The phantom data sheet provides the physical dimensions of the phantom, and the circumference of the phantom measures 86cm. Whilst this information is useful, it would be

more clinically relevant to describe the phantom in terms of patient BMI, as this information is more readily available for patients and gives an indication as to how large the patient is likely to be (patient weight is not necessarily a good indicator of patient habitus).

Studies from Japan [145], [229] state the NEMA body phantom represents a 60kg patient, but do not specify height or BMI. Some studies have adapted the phantom in order to represent larger patients: e.g. patients larger than 70kg [145] or patients with BMI > 30 [19]. These studies therefore implicitly suggest the phantom represents a patient of less than 70kg, or a patient with BMI less than 30.

According to the National Office for Statistics in 2010 [230], the average BMI in the UK for both males and females was approximately 27. It would therefore be useful to determine if the NEMA body phantom can be reliably used to represent PET imaging of an average patient. To the author's knowledge, there is no published work explicitly comparing the NEMA body phantom to a range of patient sizes in order to determine which patient sizes can be usefully represented by the phantom.

10.2 Aims

The aim of this chapter is to determine how well the NEMA IEC body phantom represents patient liver imaging. More specifically, the chapter aims to determine how applicable the phantom results of Chapter 9 are likely to be to clinical liver imaging. Phantom images will be compared with liver images from patients with varying BMI. The following comparisons will be made in this chapter:

1. Physical dimensions in the transaxial plane
2. Mean activity concentration in liver versus phantom background chamber
3. Image noise in liver versus phantom background chamber

10.3 Materials and Methods

10.3.1 Patient Selection

Patients reported as having normal liver uptake and a range of different weights/BMIs at the time of their PET scans were selected for this study. Weights ranged from 41kg to 141kg, with BMI ranging from 19.5 to 52. Ten patients were identified for each of the three following BMI categories:

- BMI $\leq$ 25
- 25 < BMI $\leq$ 35
- BMI > 35

10.3.2 Image Acquisition and Reconstruction Protocol

The same preparation, acquisition and scanning protocol was followed for all thirty patients. Approximately 400MBq 18F-FDG was administered following a 6-hour fast. Once residual activities were taken into account, actual injected activities ranged from 344MBq to 419MBq (average 375MBq). PET imaging was performed following a 60-minute uptake period using 4-minute acquisition beds. Patients were scanned with their arms above their heads when possible (14 patients). When this was not possible, arms were placed by their sides, and were therefore in the liver FOV (16 patients).

The phantom acquisition used in this study was the same acquisition used in Part 1 of Chapter 9: activity concentration at scan time was 5,346Bq/ml in the background chamber and 24,830Bq/ml in the spheres (4.64:1 ratio). Images were acquired using 4-minute beds (two bed positions were acquired with the spheres in the FOV centre).

All patient and phantom images were reconstructed using the parameters suggested by GEMS for clinical imaging: TOF and PSF were both applied with 54 effective iterations, 192 matrix, standard z-axis filter and 4mm Gaussian filter.

10.3.3 Image Analysis

A single transverse slice showing the largest cross section of the liver was selected for patient analysis, as shown in Figure 10.1 (top row). CT images were used to measure patients' dimensions in the horizontal (right-left) and vertical (anterior-posterior) directions. Arms were not included in these measurements if they were in the FOV. A 5cm ROI was placed on the liver on the PET images. This was used to record the mean and standard deviation of the voxels. These were used to calculate COV (as per Equation 3.2). The same measurements were made using the body phantom (also shown in Figure 10.1, bottom row).

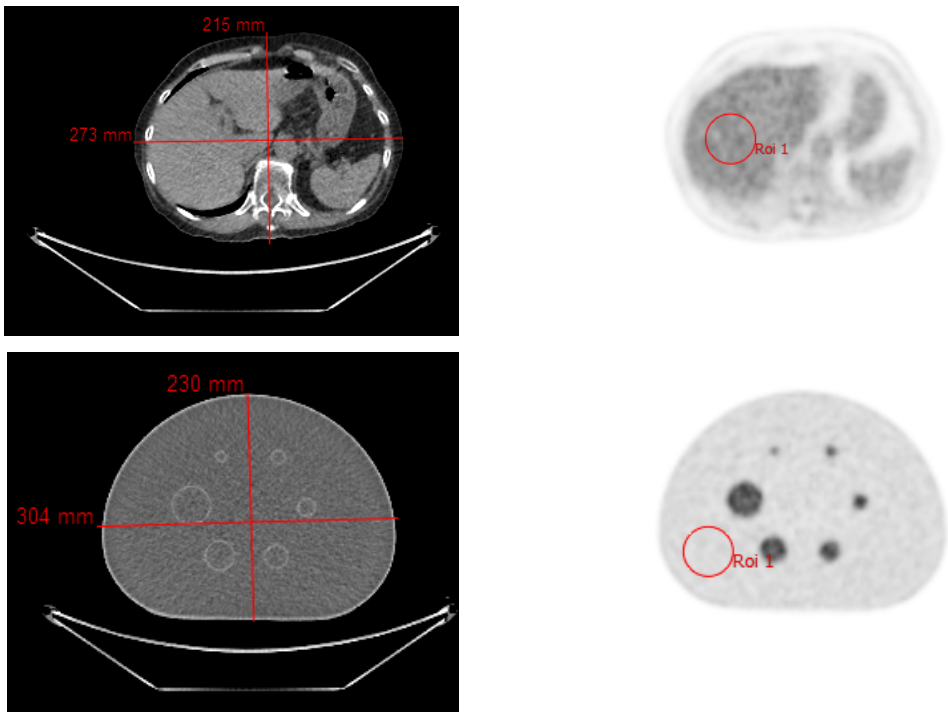


Figure 10.1: Patient (top row) and phantom (bottom row) image measurements Physical dimensions were measured on CT data (left). Mean activity concentrations and image noise were measured on PET data (right).

10.4 Results

10.4.1 Physical Dimensions

The measured physical dimensions of the phantom matched its published dimensions: 300mm at the maximum horizontal width (right/left direction on images) and 230mm at the maximum vertical height (anterior/posterior direction on images).

A scatter plot of the patient torso measurements is shown in Figure 10.2. The x-axis shows the measurement in the right/left plane while the y-axis shows the measurement in the anterior/posterior plane. Each data point represents a single patient, colour coded by BMI category. Dashed lines represent the phantom dimensions, while the shaded boxes represent ± 10% of the phantom dimensions.

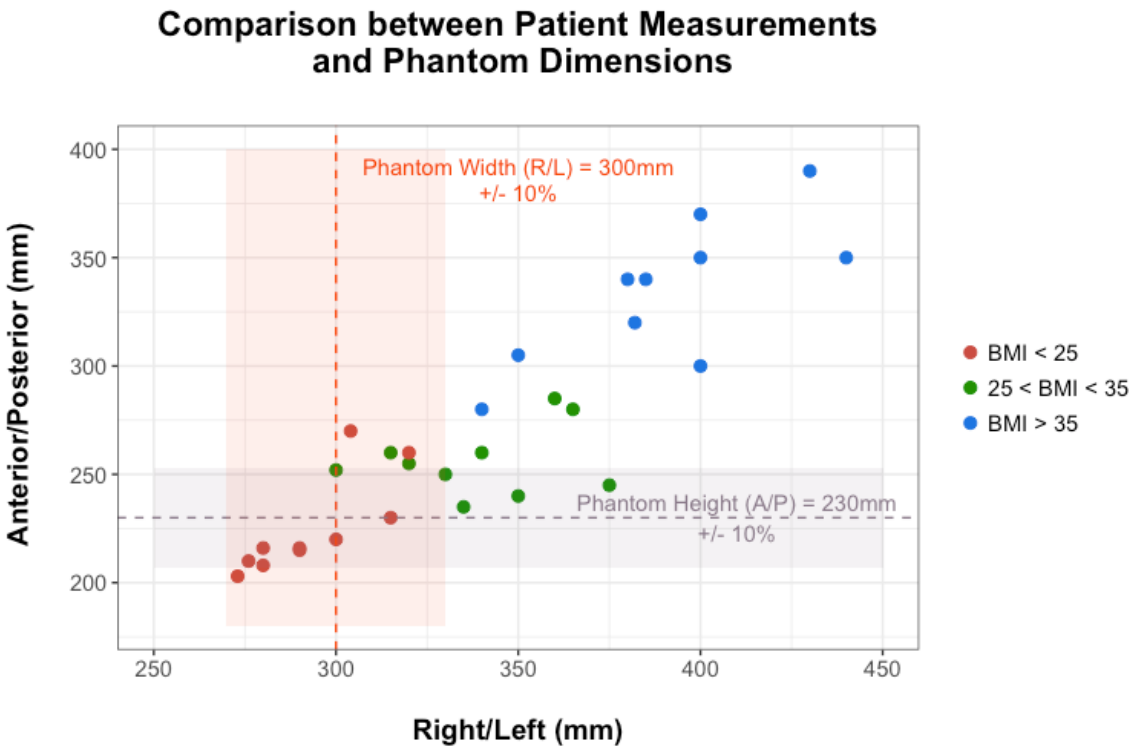


Figure 10.2: Comparison between patient measurements and phantom dimensions

Dashed lines represent phantom dimensions. Shaded areas represent ± 10% of phantom dimensions.

Figure 10.2 demonstrates that nine of the thirty patients had both R/L and A/P measurements within 10% of those of the phantom (these data points lie within the intersection of the two shaded areas). A further five patients' R/L measurements (orange shaded area) and three patient's A/P measurements (purple shaded area) were within 10% of the corresponding phantom dimension. The remaining thirteen patients' measurements fell outside ± 10% of the phantom dimensions.

With respect to BMI, Figure 10.2 demonstrates all ten patients with BMI ≤ 25 had at least one measurement within 10% of the phantom dimensions. Conversely, none of the patients with BMI > 35 were within 10% of either of the phantom dimensions. Seven of the 25 < BMI ≤ 35 patients had at least one measurement within 10% of the phantom dimensions.

10.4.2 Mean Activity Concentrations

Figure 10.3 plots mean patient liver activity concentrations against injected activities (corrected for residual syringe measurements). The dashed reference line corresponds to the phantom's background activity concentration measurement (5,569Bq/ml), whilst the grey shaded area corresponds to the phantom measurement ± 10%.

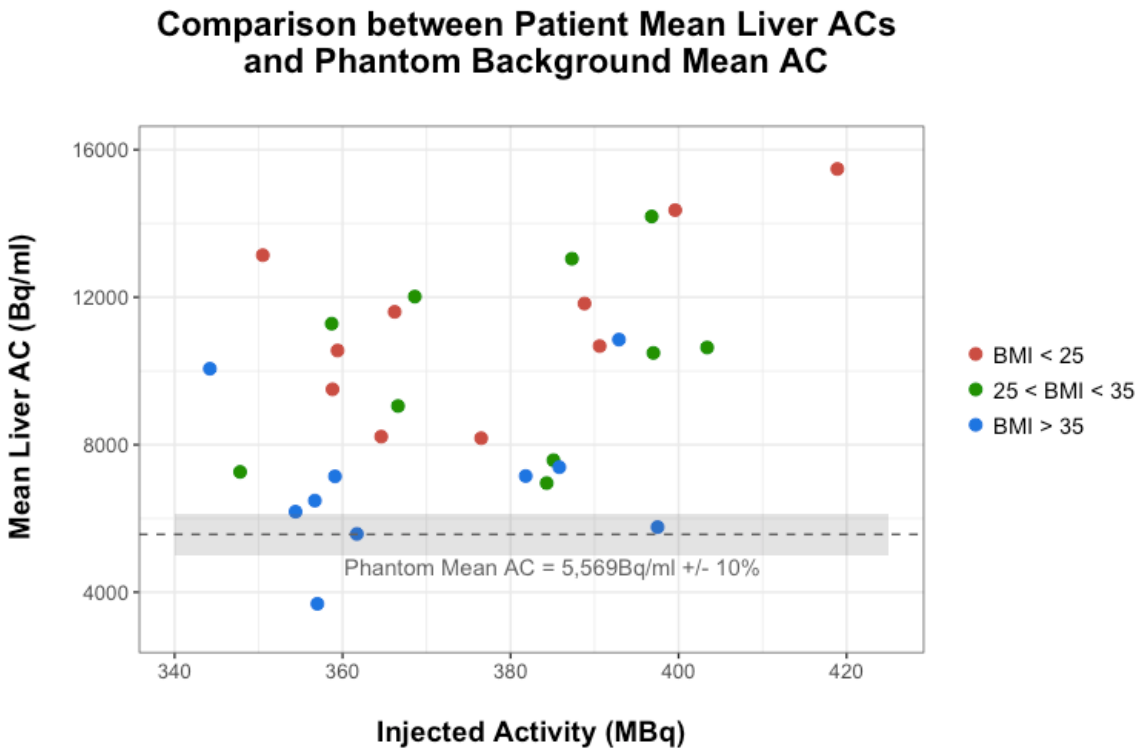


Figure 10.3: Patient liver activity concentrations versus injected activity
Dashed line represents phantom background mean activity concentration. Shaded region represents ± 10% of phantom measurement.

The choice of phantom background activity used in this thesis, previously discussed in Chapter 4, was designed to be conservative with respect to typical patient liver activity concentrations when 400MBq ^{18}F-FDG is administered and a sixty-minute uptake period is observed prior to imaging. Figure 10.3 demonstrates that twenty-nine of the thirty patients had a mean liver activity concentration equal to or greater than that of the phantom. Nine patients (all BMI ≤ 35) had mean liver activity concentrations more than double that of the phantom. The only patient in Figure 10.3 with a liver activity concentration lower than that of the phantom was a particularly heavy patient (136kg, BMI = 41.5) who received a lower than intended injection activity (357MBq).

As the activity concentration results are small in number and in multiple, unmatched groups, a Kruskal Wallis test was applied to determine if there were any statistically significant differences between the three BMI categories. The resultant p-value of 0.002 indicated a significant difference was present. Pairwise Wilcoxon rank sum tests with Bonferroni multiple comparison corrections were therefore performed on the three BMI categories. This test demonstrated no significant difference between the two lower BMI categories ($p = 0.945$), whereas the 'BMI > 35' category was significantly different to both 'BMI ≤ 25' ($p = 0.002$) and '25 < BMI ≤ 35' ($p = 0.016$) categories.

10.4.3 Image Noise

Figure 10.4 plots patient liver noise against patient BMI. Data is presented as two groups by arm position, with linear trendlines included. The dashed reference line corresponds to the phantom's background noise (10.4%), whilst the grey shaded area corresponds to the phantom noise measurement ± 10%.

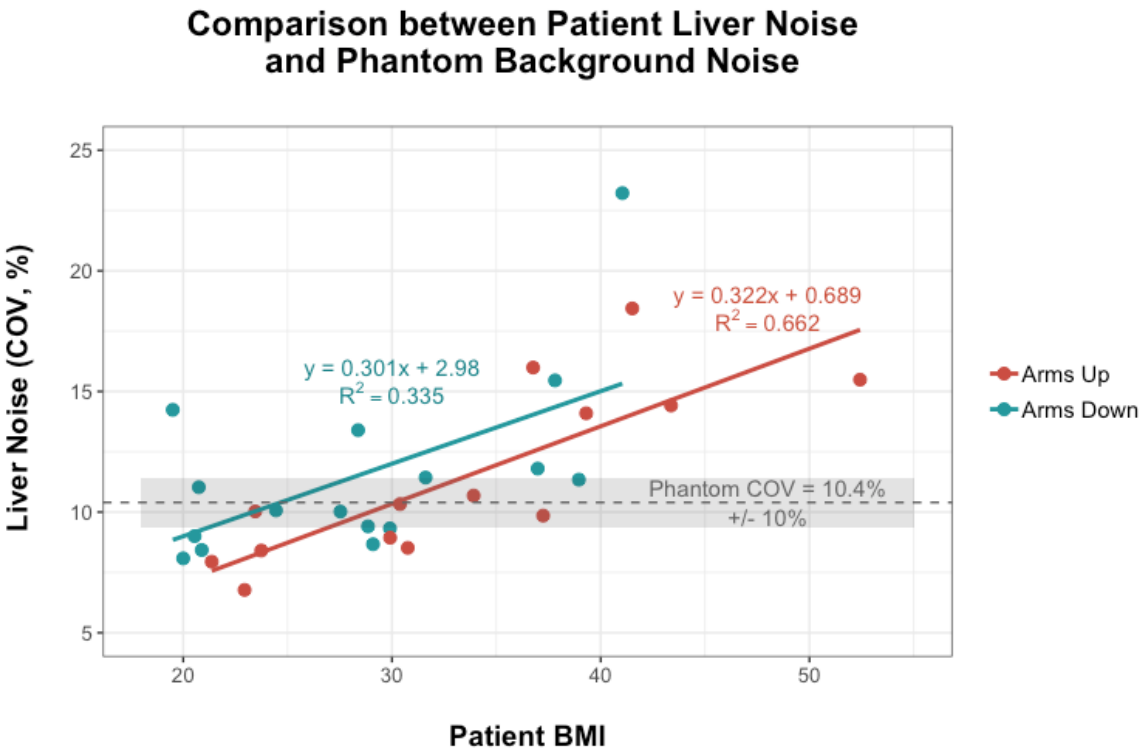


Figure 10.4: Patient liver noise versus patient BMI

Dashed line represents noise in phantom background. Shaded region represents ± 10% of phantom noise measurement. Linear trendlines are presented for different arm positions.

Figure 10.4 demonstrates liver noise tends to increase as patient BMI increases for both arm positions, as one would expect: linear regression significance tests yielded p = 0.0004 for 'Arms Up' and p = 0.0189 for 'Arms Down'. Furthermore, the linear trendlines suggest that patients scanned with their arms up (and therefore out of the liver FOV) demonstrate lower liver noise than patients scanned with their arms down by their sides, which is also as one would expect; however, there was no statistically significant difference in noise between the 'arms up' and 'arms down' result groups (Wilcoxon rank sum test p = 0.854).

Figure 10.4 also demonstrates that sixteen of the thirty patients had noise measurements lower than the phantom, with a further four patients within 10% of the phantom noise measurement. Of the ten patients who exceeded the phantom noise by more than 10%, eight had BMI > 35. The remaining two patients were scanned with their arms down by their sides.

Seven of the eight patients with BMI ≤ 35 who were scanned with their arms up had a noise measurement less than that of the phantom, while the ninth patient's noise measurement was within 10% of the phantom noise.

A Kruskal Wallis test indicated significant differences in noise measurements between the three BMI categories, regardless of arm position ($p = 0.0006$). Pairwise Wilcoxon rank sum (Bonferroni) tests demonstrated no significant difference between the two lower BMI categories ($p = 0.653$), whereas the 'BMI > 35' category was significantly different to both 'BMI ≤ 25' ($p = 0.001$) and '25 < BMI ≤ 35' ($p = 0.002$) categories.

10.5 Discussion

The physical comparisons between the phantom and patients demonstrate the phantom's limitations as a liver surrogate. The dimensions of the phantom appear to be a reasonable match for patient torsos when BMI ≤ 25 and may also be considered representative of patients with BMI between 25 and 35. However, the phantom appears to be too small to represent patients with BMI > 35. One would therefore expect that, with the use of appropriate activity concentrations and counting statistics, the NEMA body phantom is a useful representation of patients with BMI ≤ 25, and may also be useful for patients with BMI between 25 and 35. This finding appears to be consistent with studies that adapted the phantom for larger patients [19], [145]. However, the liver occupies approximately 25% of the patients' torso images, whereas 100% of the body phantom has been taken to represent the liver in this thesis. Additionally, the phantom on its own does not simulate the effects of a patient with their arms by their sides or having breast tissue in the FOV. Comparisons of PET data measurements were therefore required to determine how the phantom's physical limitations affected the clinical relevance of the phantom analysis undertaken in this thesis so far.

The activity concentration in the phantom's background chamber was deliberately chosen to be lower than that typically observed in patient livers in order to provide conservative counting statistics (as discussed in Chapter 4). Figure 10.3 demonstrated the background activity concentration used in this chapter was particularly conservative for patients with BMI ≤ 35. Larger patients (BMI > 35) had lower liver activity concentrations than smaller patients, as one would expect when patients of all sizes are injected with the same activity. There were no statistically significant differences in activity concentrations between the 'BMI ≤ 25' and '25 < BMI ≤ 35' patients; however, the 'BMI > 35' activity concentrations were significantly lower than both other BMI categories. The phantom therefore did not provide a

conservative estimate of larger patients' liver counting statistics; one patient with BMI > 35 demonstrated a lower activity concentration than the phantom.

Finally, patient noise measurements were compared with those from the phantom. The conservative choice of phantom activity concentration and hence counting statistics would ideally result in the phantom having greater noise levels than the patient population. This was the case for all but one of the eight patients with BMI < 35 who were scanned with their arms out of the liver FOV (the eighth patient's noise was within 10% of the phantom). Furthermore, only two of the 12 patients with BMI $\leq$ 35 and arms down had noise levels that exceeded phantom noise by more than 10%. There were no statistically significant differences in noise levels between the 'BMI $\leq$ 25' and '25 < BMI $\leq$ 35' patients; however, the 'BMI > 35' noise results were significantly greater than both other BMI categories.

One may have expected the conservative choice of phantom activity concentration to provide a more conservative noise estimate than that demonstrated in Figure 10.4; however, the phantom noise measurement appears to be more representative of the clinical scenario than the phantom activity concentration. The reason for this discrepancy is likely to be caused by the inherent difference in uptake patterns between a water phantom and a patient liver. The phantom's background chamber is filled only with liquid and results in homogenous activity concentrations throughout the volume, provided the ^{18}F-FDG is sufficiently mixed: there is no tissue-like substance placed inside the phantom chamber to mimic the heterogenous composition (and therefore heterogenous ^{18}F-FDG uptake) of the human liver. This means that if the same count density was used for both phantom and patient liver imaging, the resultant images would likely produce similar activity concentrations, but the liver images would have greater noise measurements than the phantom images. This chapter therefore demonstrates a conservative choice of phantom background activity is required when using the NEMA body phantom to represent liver imaging, in order to produce clinically relevant noise measurements.

As one would expect, this section also demonstrates that noise increased as BMI increased, as all patients studied in this chapter were prepared and scanned using the same protocol regardless of size. It should be noted that weight-based ^{18}F-FDG administration and scanning protocols can be used to produce more consistent noise results amongst patients of different sizes [14], [179], [186].

10.6 Conclusions

The aim of this chapter was to determine how well the NEMA body phantom represents the patient liver, and therefore predict how applicable the reconstruction recommendations made by Chapter 9 would be to clinical liver imaging. Although the phantom has physical limitations as a liver surrogate, it appears to provide a good representation, in terms of image noise, for livers in patients with BMI $\leq$ 25. Furthermore, the phantom also appears to provide a reasonable representation of livers in patients with BMI $\leq$ 35, particularly when the patients' arms are not placed in the liver FOV. As the average UK BMI is approximately 27, the phantom is therefore expected to provide good representation of the average UK patient liver when a suitably conservative activity concentration is used.

The phantom work undertaken in Chapter 9 is therefore expected to provide reasonable predictions for image quality, and hence lesion detection, in patients with BMI < 35. It may not be as accurate for patients with BMI > 35; the reconstruction strategies recommended by Chapter 9 are unlikely to be optimal for these larger patients. The lesion detection performance of the recommended reconstructions will be tested using patient data in the following chapter.

The BMI range represented by the phantom could be determined more accurately by repeating this work with a greater number of patients and smaller BMI category ranges. The NEMA body phantom could be improved in future work by including scatter material round the phantom to simulate patients with higher BMIs, in a similar manner to studies that adapted the phantom for larger patients [19], [145]. Furthermore, the use of phantom 'arms' and 'breasts' could be used to improve the phantom's relevance to different clinical imaging scenarios.

Chapter 11 : Qualitative Observer Study of Patient Data

In Chapter 9, multiple reconstructions of a body phantom were performed using different combinations of reconstruction parameters. Qualitative and quantitative analyses were undertaken to identify which combinations of reconstruction parameters optimised detection of the smallest sphere within the phantom. An informal observer study, combined with quantitative analysis, concluded that two specific combinations of reconstruction parameters in particular merited further evaluation in a patient study.

Furthermore, Chapter 10 concluded that the body phantom was a good representation of patients with BMI < 25, in terms of image noise. The phantom was also found to be a reasonable representation of patients with BMI between 25 and 35. It is therefore expected that the body phantom work undertaken in Chapter 9 should provide reasonable predictions for image quality in patients with BMI < 35.

A major limitation of Chapter 9's observer study was the spheres' fixed sizes, positions and sphere-to-background ratio. Furthermore, the activity concentration within the phantom's background chamber is more homogenous than liver background activity in clinical ^{18}F-FDG scans; it is therefore less challenging to identify a small lesion within the phantom's background than it is within a liver. A patient study is therefore required to assess reconstruction performance in a clinically relevant manner.

11.1 Introduction

This chapter conducts a formal human observer study of patient data using the two reconstruction methods identified by Chapter 9's phantom analyses. The clinical

reconstruction protocol suggested by GEMS is also included. The inclusion of a fourth reconstruction method, designed to bridge the gap between the GEMS reconstruction and Chapter 9's recommended reconstructions, is also justified later in this chapter. The results of this observer study will be used to make recommendations for clinical liver image reconstruction when using the GEMS Discovery 690 PET-CT system, thus fulfilling the secondary aim of this thesis.

11.1.1 Review of the Literature

Receiver Operator Characteristic (ROC) analysis is a well-established method for statistically quantifying differences between the detection performance of observers for different imaging modalities, where the known truth is available [32], [231]. Observers indicate whether each case is normal or diseased, and state how confident they are in their decision using a rating scale. Observer results are then compared to the known truth for each patient. Figure 11.1 illustrates the principles of ROC plots: True Positive Fraction (TPF) is plotted against False Positive Fraction (FPF) as the decision criterion is varied from strict (high confidence) to lenient (lower confidence). The Area under the Curve (AUC) is an objective figure of merit ($0 \leq AUC \leq 1$) and is the probability that an abnormal (diseased) image will be rated higher than a normal (non-diseased) image. Any diagnostic test with an $AUC \leq 0.5$ is a poor test, as any results occur by chance [231]: this is represented by the 'guessing line' and the grey shaded area in Figure 11.1.

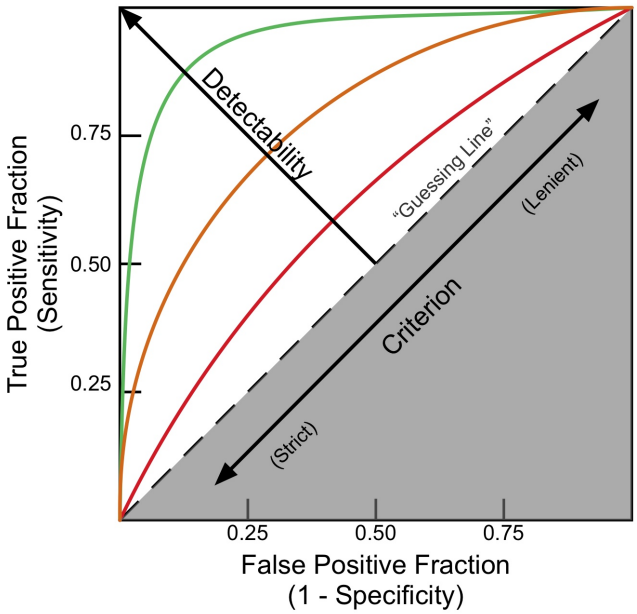


Figure 11.1: Principles of Receiver Operator Characteristic (ROC) plots

Green curve represents a good diagnostic test, orange represents a fair diagnostic test, red represents a poor diagnostic test. Dashed line represents a test that relies on chance. Plot adapted from [231]**.**

ROC analysis has been used in many PET imaging studies [150], [151], [180], [219], [232]–[236]; however, it has limitations. This type of analysis can only be used for binary tasks (i.e. the image is either normal or abnormal) and does not account for the position of the identified lesion. Furthermore, it cannot use all diagnostic information potentially available from the observers as it ignores the presence of multiple lesions.

An improved version of ROC is Localisation ROC (LROC). This method informs the observer there can be at most one lesion per image and requires the observer to record the location of the detected lesion, which is then taken into consideration by the analysis. The observer also records their confidence in the decision. Plots of TPF versus FPF, and AUC, are produced in a similar manner to ROC analysis. Although LROC analysis is a widely accepted technique for measuring the performance of imaging tasks, and has been used in several PET studies comparing different reconstruction methods [8], [110], [157], [213], [215], [217], [218], [220], it remains limited to the use of one lesion per image/patient. In PET clinical liver imaging, the 'truth' within the liver could be 'multifocal'; i.e. there could be more than one lesion in the liver, and the detection of all lesions could be crucial to patient management (curative surgical resection versus palliative treatment).

A further variant of ROC analysis, known as Free Response ROC (FROC), allows a reporting clinician to mark multiple suspicious locations on patient images and rate their confidence in each marked location being a lesion [237]. These 'mark-rating' pairs are used to compare different imaging methods or modalities (for example, different methods of PET reconstruction). The analysis requires the 'truth' to be known to the study coordinator, while the observers should be blinded: they are given no prior information regarding the number of lesions in the image. FROC analysis enables more precise evaluation of imaging systems as the location of each marked lesion is accounted for in the analysis [238].

FROC curves differ from ROC curves: the fraction of lesions located correctly (LLF) ($0 \leq$ LLF ≤ 1) are plotted against the number of false locations per image (NLF) ($0 \leq$ NLF $\leq \infty$). The resultant FROC plot visualises the comparative performance for each reconstruction/observer and provides an indication of how fully the observers used the reporting scale. For example, an observer who is more likely to report low confidence lesions will produce a FROC curve that extends further to the right than an observer who only reports high confidence lesions. This is illustrated by a simple example in Figure 11.2 (a), based on two fictional observers who applied different confidence thresholds when reporting lesions. Observer 1 applied a lower minimum confidence threshold than Observer 2. As a result, Observer 1 was able to identify more true lesions than Observer 2 (71% versus 60%), but at the expense of detecting more false positives (4 versus none, over a

total of 30 images). Observer 2's FROC curve is a simple vertical line at x = 0 (no false positives) which ends at y = 0.6 (60% of lesions identified). Observer 1's curve is more complex; it terminates at x = 0.13 (4 false positives divided by 30 images) and y = 0.71 (71% of lesions identified). The various points on the curve between the origin and the terminal point represent the cumulative fraction of lesions identified at progressively lower confidence levels (moving from left to right).

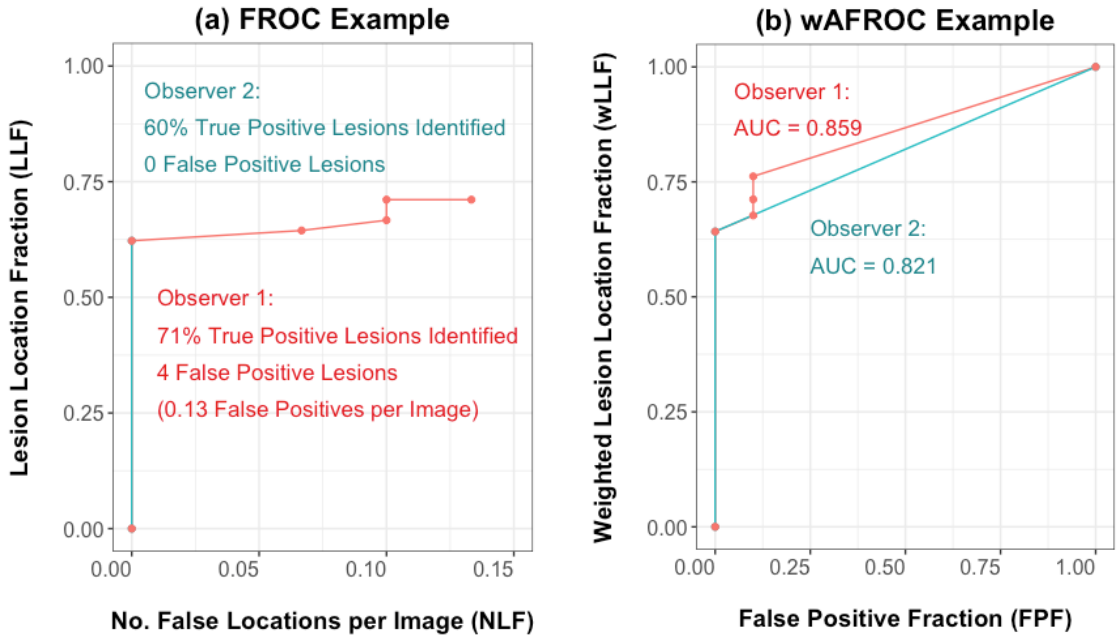


Figure 11.2: Example illustration of (a) FROC and (b) wAFROC curves
Observer 1 detected more challenging lesions with higher weightings by applying a lower confidence threshold.

FROC curves have several limitations. They don't account for correctly unmarked non-disease cases. Furthermore, it is difficult to use them to produce a meaningful quantitative figure of merit (e.g. AUC) as they are unconstrained on the x-axis. Alternative FROC (AFROC) plots are therefore used to perform AUC analyses. The most widely used method for statistically analysing such data is Jackknife AFROC (JAFROC), which has undergone extensive validation [239]. A full explanation of such analysis is beyond the scope of this thesis; briefly, AFROC plots the fraction of true lesions identified correctly ($0 \leq LLF \leq 1$) against the inferred false positive fraction ($0 \leq FPF \leq 1$). Weighted AFROC (wAFROC) analyses allow weightings to be assigned to individual lesions in the study; for example, smaller lesions which are challenging to identify may be assigned a greater weighting than larger lesions which are less challenging to identify. AUCs generated from wAFROC analyses are the recommended figure of merit for FROC style observer studies [240]. Figure 11.2 (b) demonstrates that Observer 1 produced a greater AUC than Observer 2 once lesion weightings and confidence levels were fully accounted for. FROC-style analyses have been

performed in several PET observer studies [241]–[244]; however, none of these studies assessed liver lesion detection using TOF and PSF reconstructions.

All ROC variants require the 'truth' to be known. The PET studies considered in this literature review achieved this using both phantom and patient studies. Phantom studies, by definition, have a known truth to compare against human observer assessments. However, knowledge of the truth within patient studies can be more challenging. Several publications used sinogram manipulation to combine healthy patient data with separately acquired lesion data [8], [110], [241], [242], [245]. The number and location of lesions were therefore known to the study supervisors. Other studies used diseased patient data in their observer studies: appropriate correlative imaging (CT/MRI), clinical follow-up and/or histopathology were used as gold standards [243], [246], [247].

11.2 Aims

This patient observer study aims to evaluate and compare the lesion detection capabilities of four different combinations of reconstruction parameters using a liver lesion detection assessment. Observers will attempt to correctly determine both the presence and absence of liver lesions using a group of anonymised patient studies. Observers will also assess overall image quality and state their confidence in their normal/abnormal liver diagnosis.

The aim of this chapter is to recommend one of the four reconstruction methods for clinical liver imaging in order to optimise small lesion detection within the liver. Results will be discussed with reference to patient BMI, as this is likely to affect the observers' assessment of the images.

Furthermore, this chapter aims to establish a generic methodology for conducting clinical observer studies for a wider range of clinical applications.

11.3 Materials and Methods

11.3.1 Patient Selection

PET reporting clinicians were asked to record details of potential patients of interest for a liver reconstruction study, scanned on the GEMS Discovery 690 PET-CT system, over a period of 12 months in 2013/2014. Fifty-six patients were identified in this manner, the majority of whom had a history of colorectal cancer. Sinogram data for these patients were stored to enable retrospective reconstructions at a later date. A group of patients reported to have healthy livers over the same time period were also identified for this study.

Patient records were checked for correlative imaging, resections, pathology results and follow-up imaging. MRI is considered to be the gold standard for the evaluation of hepatic lesions [98], while pathology can provide definitive answers to the nature of lesions (where available). 36 of the 56 patients identified by the clinicians did not have an appropriate gold standard available to verify the true/false nature of identified lesions and were therefore eliminated from the study.

All remaining twenty patients had a history of colorectal cancer. As discussed in Chapter 1, up to 70% of patients with colorectal cancer will develop liver metastases at some point [248]. The observer study was therefore designed to have approximately 70% diseased liver patients and 30% healthy liver patients. Ten patients with no liver disease (confirmed by follow-up) were added to the study, creating a patient group with 67% disease prevalence.

Patient weights ranged from 42kg → 115kg (average 75.8kg), with a BMI range 16.4 → 43.8 (average 27.9). Patients were split into the same three BMI categories used in Chapter 10: BMI $\leq$ 25 (9 patients), 25 < BMI $\leq$ 35 (18 patients) and BMI > 35 (3 patients). The numbers of patients in the BMI > 35 category is particularly low; results are included for completeness but cannot be used to draw any meaningful conclusions.

A ‘true’ lesion database for all thirty patients was created with the assistance of an experienced PET reporting clinician and the available gold standard data. An MRI reporting clinician was consulted for some of the more challenging cases. The location of each confirmed lesion was recorded using Hermes Hybrid Viewer software. The centre of each lesion was selected with a triangulation tool and the transverse, coronal and sagittal locations were recorded. Three patients had additional lesions confirmed by MRI that were not visible on the original clinical PET reconstruction: in these cases, the lesion location on the PET data was estimated using the MRI data. In total, 45 ‘true’ lesions were identified across all twenty diseased patients. The ten ‘normal’ patients were confirmed to have no liver disease.

11.3.2 Acquisition Protocol

The same preparation, acquisition and scanning protocol was followed for all thirty patients. Approximately 400MBq ^{18}F-FDG was administered following a 6-hour fast. Once residual activities were taken into account, injected activities ranged from 303MBq to 425MBq (average 369MBq). PET imaging was performed following a 60-minute uptake period using 4-minute acquisition frames. Patients were scanned with their arms above their heads when

possible (n = 17). When this was not possible, arms were placed by their sides, and were therefore in the liver FOV (n = 13).

11.3.3 Reconstruction Protocols

This chapter compares four different combinations of reconstruction parameters, as detailed in Table 11.1.

Reconstruction Name	Reconstruction Parameters				
	Method	**Effective Iterations**	**Gaussian Filter (FWHM)**	**Z-axis Filter**	**Matrix**
'GE Clin'	PSF+TOF	54	4mm	Standard	192*192
'Sharp 1'	PSF+TOF	54	0mm	None	256*256
'Sharp 2'	PSF+TOF	108	0mm	None	256*256
'Sharp 3'	PSF+TOF	180	2.5mm	None	256*256

Table 11.1: Patient observer study reconstruction parameters

The following summarises the justification for including each of these four reconstructions in the patient observer study:

- **'GE Clin'**: These are the reconstruction parameters suggested by GEMS for routine clinical whole-body imaging. This reconstruction has the lowest resolution and the lowest noise of the four reconstructions. Local clinicians currently use this reconstruction when reporting PET images.
- **'Sharp 1'**: This is a 'sharper' version of the GEMS suggested reconstruction: filtering is removed, and the matrix size is increased. The number of effective iterations remains the same as the GEMS suggested parameters. This reconstruction was chosen to bridge the gap in noise/resolution between 'GE Clin' and 'Sharp 2'.
- **'Sharp 2'**: This reconstruction was shown to provide the optimal quantitative trade-off between small sphere $HCRC_{mean}$ and image noise in Chapter 9. This reconstruction was also preferred qualitatively by three of the four Chapter 9 observers in terms of small sphere detection.
- **'Sharp 3'**: This reconstruction was preferred qualitatively by one of the four observers in Chapter 9 in terms of small sphere detection. It has the highest resolution and the greatest noise of the four reconstructions.

All thirty patients were reconstructed using each of the four reconstructions outlined in Table 11.1, producing a total of 120 image volumes.

11.3.4 Observer Evaluation of Reconstructed Patient Images

The patient observer study was performed by two experienced PET-CT reporting clinicians using Hermes Hybrid Viewer software. All 120 patient image volumes were anonymised such that observers were blinded to both patient identifiers and the reconstruction method. Image volumes were divided into four batches of 30. Batches were designed such that no patient appeared twice in any batch, and each batch contained an approximately even mix of all four reconstruction methods. Observers were asked to wait at least one week after completing one batch before beginning the next batch.

Observers were asked to record the following data, using only the liver images:

1. **Image Quality:** Observers were asked to rate the quality of the images, in their personal opinion, using a 5-point ordinal scale: '1' signified 'very poor' image quality and '5' signified 'excellent' image quality.
2. **Overall Diagnosis: Normal or Abnormal Liver:** Observers were asked to state if the liver was normal or abnormal. They were also asked to state how confident they were with their diagnoses as a percentage, with 100% signifying absolute certainty.
3. **Location of Lesions:** Locations of any identified lesions were 'marked' by triangulating to the lesion centre and recording the transverse, sagittal and coronal image slice numbers. A percentage confidence rating was also recorded for each lesion, with 100% signifying absolute certainty a lesion was present at the identified location. Data for this section was therefore collected in 'mark-rating pairs'.

11.3.5 Analysis of Observer Study

11.3.5.1 Image Quality

The median image quality score for each reconstruction method was calculated as a comparative measure. As the data are non-parametric and in matched groups (reconstructions are matched by patient), a Friedman's test was performed to determine if there were any statistically significant differences between the scores for the four different reconstruction methods. In the event that the Friedman's test found there were significant differences, pair-wise Wilcoxon's signed rank tests with Bonferroni multiple comparison corrections were performed to determine where the significant differences lay. Agreement between both observers' image quality scores was assessed using Cohen's Kappa test. Table 11.2 outlines how the resultant Kappa coefficient, κ, should be interpreted. Each observers' image quality results were assessed separately. The effects of patient BMI upon image quality results were also assessed.

κ	**Interpretation**
< 0	Poor Agreement
0.01 → 0.20	Slight Agreement
0.21 → 0.40	Fair Agreement
0.41 → 0.60	Moderate Agreement
0.61 → 0.80	Substantial Agreement
0.81 → 1.00	Almost Perfect Agreement

Table 11.2: Kappa agreement measures for categorical data [249]

11.3.5.2 Overall Diagnosis

Average confidence in reporting each reconstruction method were calculated with 95% confidence intervals. Any statistical differences between reconstruction methods were confirmed using Friedman ± Wilcoxon/Bonferroni tests. Each observers' confidence results were assessed separately. The effects of patient BMI upon diagnostic confidence were also assessed.

Cohen's Kappa test was again used to determine observer agreement for the normal/abnormal diagnosis. The overall normal/abnormal diagnosis results were compared with the gold standard data to determine the following metrics for each of the four reconstructions:

- Sensitivity: % abnormal patients identified correctly
- Specificity: % normal patients identified correctly
- Diagnostic Accuracy: % correct overall diagnoses
- Positive Predictive Value (PPV): % abnormal diagnoses correct
- Negative Predictive Value (NPV): % normal diagnoses correct

The effects of patient BMI upon the above metrics were also assessed.

11.3.5.3 Location of Lesions

A FROC/wAFROC analysis was performed on the mark-rating pairs. Each potential lesion identified in the observer study was compared to the 'true' lesion database and categorised as being either a True Positive or a False Positive. Marked lesions were classed as True Positives if their location coincided with that of the 'true' lesion. All other marked lesions were classed as False Positives. The 'RJafroc' package for 'R' was then used to produce the following analyses:

- FROC curves for each reconstruction method (split by observer)
- wAFROC curves for each reconstruction method (split by observer)
    - smaller lesions were assigned greater weightings than larger lesions, as they are more challenging to detect
- AUC figures of merit for each reconstruction method.

11.4 Results

Figure 11.3 shows reconstructed images for three example patients, one from each BMI category.

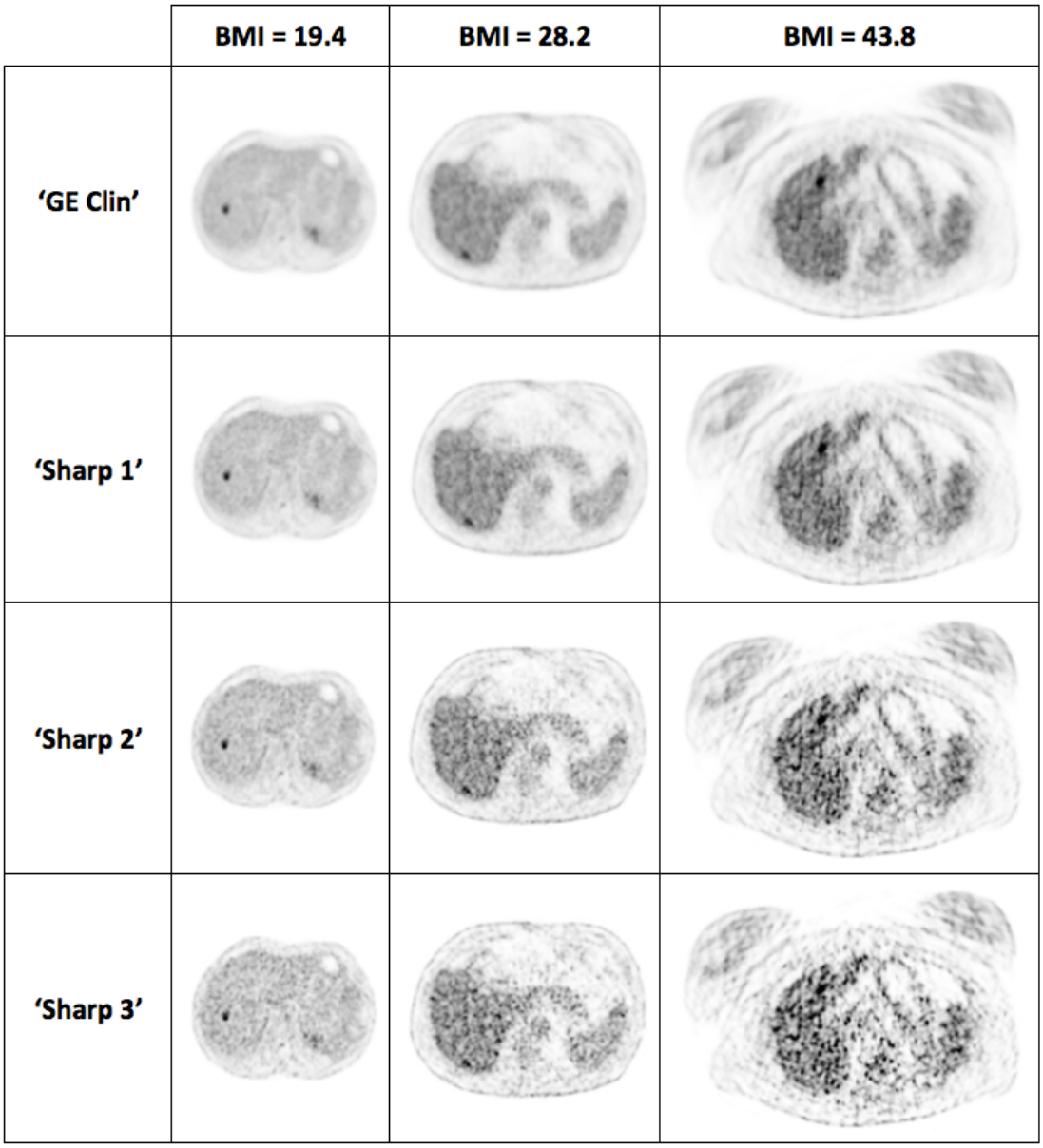


Figure 11.3: Image quality comparisons for different patient BMI categories

The effects of BMI on image quality are demonstrated by the patient images in Figure 11.3.

All three patients demonstrate increased image noise as reconstruction sharpness is increased. Lesions also appear to demonstrate increased intensity. The smallest patient's (BMI = 19.4) image quality appears to be reasonable when the sharpest reconstruction is used: the lesion is easily identifiable within the liver. However, the largest patient's (BMI = 43.8) image quality is poor when both 'Sharp 2' and 'Sharp 3' are used: it is difficult to visualise the lesion in either of these reconstructions. One would therefore expect observers to prefer the smoother reconstructions ('GE Clin' and 'Sharp 1') for the larger patients. The sharper reconstructions may demonstrate increased lesion detection for smaller patients.

11.4.1 Image Quality Results

Table 11.3 shows the median image quality scores for each of the reconstruction methods.

	'GE Clin'	**'Sharp 1'**	**'Sharp 2'**	**'Sharp 3'**
Median Image Quality Scores:	4	3	2	2

Table 11.3: Overall median image quality scores

The Friedman test result returned a p-value of <0.001, indicating there are statistically significant differences between the four reconstruction methods. Pair-wise Wilcoxon signed ranks (Bonferroni) tests were then performed between each possible pair of reconstructions. The p-value in each case was <0.01, indicating each of the four reconstruction methods produces statistically significantly different results to each of the others (at the 5% significance level). The median results may therefore be viewed as ranking the image quality; 'GE Clin' was the preferred reconstruction, while 'Sharp 3' was the least preferred reconstruction (although the median results were the same for 'Sharp 2' and 'Sharp 3', 'Sharp 2' received more ranks above the median result).

Both observers assigned the same image quality score to 61 of the 120 reconstructions (50.8% agreement), while 114 (95.0%) of the reconstructions were scored within 1 mark of each other. The Kappa coefficient was calculated to be 0.376, which corresponds to "fair agreement" between the observers. The disparity between the two observers was therefore investigated further. Table 11.4 below compares the median image quality results and the results of pairwise Wilcoxon signed rank (Bonferroni) tests for each observer.

		‘GE Clin’	‘Sharp 1’	‘Sharp 2’	‘Sharp 3’
Median Image Quality Scores:	Observer 1	4	3	2	2
	Observer 2	4	4	2	2

Observer 1: p-values

	GE Clin	Sharp 1	Sharp 2
Sharp 1	0.5566		
Sharp 2	0.0001	< 0.0001	
Sharp 3	< 0.0001	< 0.0001	0.131

Observer 2: p-values

	GE Clin	Sharp 1	Sharp 2
Sharp 1	0.0072		
Sharp 2	< 0.0001	< 0.0001	
Sharp 3	< 0.0001	< 0.0001	0.0006

Table 11.4: Median image quality scores, per observer

Pairwise Wilcoxon signed rank test p-values are also included.

Table 11.4 demonstrates some discrepancies between the two observers. Whilst both observers produced the same median image quality scores for ‘GE Clin’, ‘Sharp 2’ and ‘Sharp 3’, ‘Sharp 1’ received a higher median score from Observer 2 than Observer 1. Statistical testing demonstrated that Observer 1 saw no significant difference between ‘GE Clin’ and ‘Sharp 1’, or between ‘Sharp 2’ and ‘Sharp 3’, while all other pairs of reconstructions produced significantly different image quality scores. In contrast, statistical testing demonstrated significant differences between Observer 2’s scores for all four reconstruction methods.

The effects of patient BMI upon image quality scores were then assessed. Table 11.5 shows the median image quality scores for each reconstruction method and BMI category. Pairwise Wilcoxon signed rank (Bonferroni) tests are also shown (no statistical testing was performed on the BMI > 35 category as the numbers are too small).

		‘GE Clin’	‘Sharp 1’	‘Sharp 2’	‘Sharp 3’
Median Image Quality Scores:	BMI ≤ 25	4	4	3	3
	25 < BMI ≤ 35	4	3	2	2
	BMI > 35	4	3	2	1

BMI ≤ 25: p-values

	GE Clin	Sharp 1	Sharp 2
Sharp 1	1.0000		
Sharp 2	0.0679	0.0069	
Sharp 3	0.0065	0.0018	0.0805

25 < BMI ≤ 35: p-values

	GE Clin	Sharp 1	Sharp 2
Sharp 1	0.1697		
Sharp 2	0.0014	0.0021	
Sharp 3	0.0015	0.0008	0.0269

Table 11.5: Median image quality scores, per BMI category

Pairwise Wilcoxon signed rank test p-values are also included.

Table 11.5 demonstrates that median image quality scores for each of the ‘Sharp’ reconstructions was dependent upon patient size. This is particularly true for ‘Sharp 3’: the median score falls from 3 (BMI ≤ 25) to 1 (BMI > 35). It is of interest that there are no statistically significant differences in image quality, at the 5% level of significance, between the ‘GE Clin’ and ‘Sharp 1’ reconstructions when patients have BMI < 35. This suggests that, for smaller patients, the noise increase associated with the increase in spatial resolution does not alter the observers’ overall perception of image quality when compared to ‘GE Clin’. Furthermore, smaller patients (BMI ≤ 25) demonstrated no significant difference between ‘GE Clin’ and ‘Sharp 2’.

11.4.2 Overall Diagnosis Results

11.4.2.1 Reporting Confidence

Figure 11.4 compares the mean confidence score for each of the four reconstructions, for each observer. 95% confidence intervals are shown. The graphs demonstrate differences between each observers’ confidence in the four reconstruction methods. Observer 1’s confidence clearly decreases as the reconstruction sharpness increases (‘GE Clin’ = 97.5%; ‘Sharp 1’ = 92.8%; ‘Sharp 2’ = 85.0%; ‘Sharp 3’ = 77.8%). Observer 2’s confidence also decreases as the reconstruction sharpness increases, but to a lesser extent than Observer 1, with all four reconstructions averaging over 90% confidence (‘GE Clin’ = 99.1%; ‘Sharp 1’ = 98.4%; ‘Sharp 2’ = 94.6%; ‘Sharp 3’ = 90.8%). The difference between ‘GE Clin’ and ‘Sharp 1’ is less than 1%.

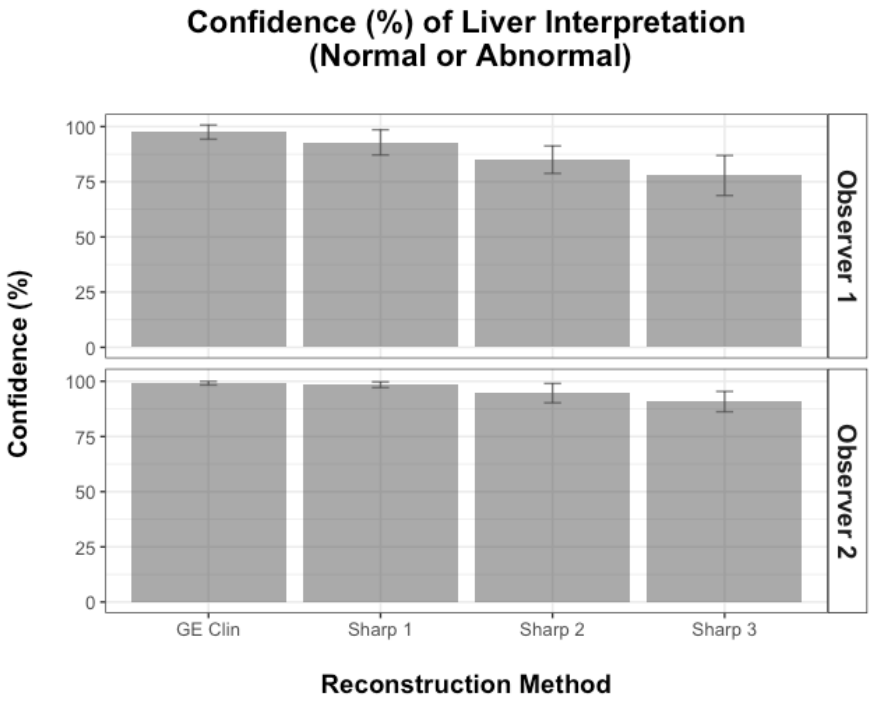


Figure 11.4: Reporting confidence, per observer

Pairwise Wilcoxon signed rank (Bonferroni) tests were applied to the individual observer confidence results. Results are shown in Table 11.6. Neither observer demonstrated any significant differences in confidence between 'GE Clin' and 'Sharp 1', or between 'Sharp 2' and 'Sharp 3'. Furthermore, Observer 2 demonstrated no significant difference between 'Sharp 1' and 'Sharp 2' (i.e. Observer 2 demonstrated no significant confidence differences in any combination of 'adjacent' reconstructions).

Observer 1: p-values

	GE Clin	Sharp 1	Sharp 2
Sharp 1	0.8959		
Sharp 2	0.0041	0.0274	
Sharp 3	0.0041	0.0377	0.3840

Observer 2: p-values

	GE Clin	Sharp 1	Sharp 2
Sharp 1	1.0000		
Sharp 2	0.0131	0.2505	
Sharp 3	0.0008	0.0080	0.1724

Table 11.6: Observer confidence significance test results

The effects of BMI upon reporting confidence are illustrated in Figure 11.5.

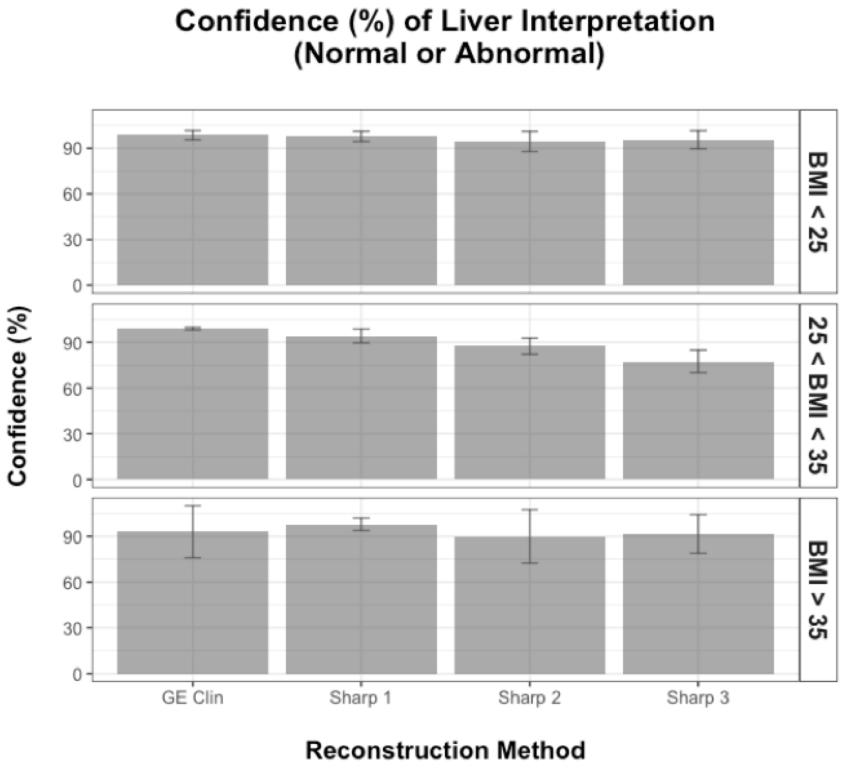


Figure 11.5: Reporting confidence, per BMI category

When BMI ≤ 25, the average confidence for all four reconstructions are similar (only 4.2% separates the highest and lowest average confidence rating) with no significant differences between any two pairs of reconstructions within this BMI category.

Conversely, the 25 < BMI ≤ 35 category demonstrates average confidence decreases as reconstruction sharpness increases. Pairs of reconstructions within this BMI category are significantly different to each other, with one exception: 'GE Clin' and 'Sharp 1' demonstrate no significant difference.

Results for BMI > 35 were not analysed for statistical significance as the numbers are too small. Figure 11.4 doesn't appear to demonstrate any meaningful data for this category.

11.4.2.2 Diagnosis Results

Observer 1 correctly diagnosed the liver as being normal or abnormal for 97 of the 120 reconstructions (80.8%), while Observer 2 was correct for 111 reconstructions (92.5%). The two observers agreed on the overall diagnosis for 104 of the 120 reconstructions (86.7%), yielding a Kappa value of 0.716 (substantial agreement).

Table 11.7 summarises the overall diagnosis results for each of the four reconstruction methods. Results are presented for each individual observer.

		'GE Clin'	'Sharp 1'	'Sharp 2'	'Sharp 3'
Sensitivity	Ob 1	95.0%	85.0%	80.0%	75.0%
	Ob 2	90.0%	90.0%	90.0%	90.0%
Specificity	Ob 1	80.0%	70.0%	60.0%	90.0%
	Ob 2	100%	90.0%	100%	100%
Diagnostic Accuracy	Ob 1	90.0%	80.0%	73.3%	80.0%
	Ob 2	93.3%	90.0%	93.3%	93.3%
PPV	Ob 1	90.5%	85.0%	80.0%	93.8%
	Ob 2	100%	94.7%	100%	100%
NPV	Ob 1	88.9%	70.0%	60.0%	64.3%
	Ob 2	83.3%	81.8%	83.3%	83.3%

Table 11.7: Summary of overall diagnosis results, per observer

Observer 2 had almost identical diagnosis results for all four reconstruction methods; the discrepancies resulted from a single false positive result for 'Sharp 1' that was not repeated for the other reconstructions. It is therefore reasonable to state that Observer 2's overall diagnosis results were independent of the reconstruction method.

In contrast, Observer 1's results appeared to have some dependency on reconstruction method, with each of the 5 metrics largely worsening as the sharpness of the reconstruction increased. The results for 'Sharp 3', however, are an exception for all metrics except sensitivity. The sharpest reconstruction produced the greatest specificity and PPV results of all four reconstructions and produced greater diagnostic accuracy and NPV results than 'Sharp 2'.

The effects of BMI on diagnostic results are summarised in Table 11.8. As there are only 3 patients in the largest BMI category (and all three were diseased patients), the metrics for BMI > 35 are of limited value (specificity, PPV and NPV are omitted from the table). It should also be noted that direct comparisons between the different weight categories should be made with caution, as their patient numbers and disease prevalence are different:

- BMI $\leq$ 25: N = 9; Normal = 3; Abnormal = 6; Prevalence = 66.7%
- 25 < BMI $\leq$ 35: N = 18; Normal = 7; Abnormal = 11; Prevalence = 61.1%
- BMI > 35: N = 3; Normal = 0; Abnormal = 3; Prevalence = 100%

		'GE Clin'	'Sharp 1'	'Sharp 2'	'Sharp 3'
Sensitivity	BMI ≤ 25	100%	100%	100%	100%
	25 < BMI ≤ 35	91.9%	86.4%	81.8%	77.3%
	BMI > 35	83.3%	66.7%	66.7%	66.7%
Specificity	BMI ≤ 25	83.3%	66.7%	83.3%	83.3%
	25 < BMI ≤ 35	92.9%	85.7%	78.6%	100%
	BMI > 35	N/A	N/A	N/A	N/A
Diagnostic Accuracy	BMI ≤ 25	94.4%	88.9%	94.4%	94.4%
	25 < BMI ≤ 35	91.7%	86.1%	80.6%	86.1%
	BMI > 35	83.3%	66.7%	66.7%	66.7%
PPV	BMI ≤ 25	92.3%	85.7%	92.3%	92.3%
	25 < BMI ≤ 35	95.2%	90.5%	85.7%	100%
	BMI > 35	N/A	N/A	N/A	N/A
NPV	BMI ≤ 25	100%	100%	100%	100%
	25 < BMI ≤ 35	86.7%	80.0%	73.3%	73.7%
	BMI > 35	N/A	N/A	N/A	N/A

Table 11.8: Summary of overall diagnosis results, per BMI category

The results for the smallest weight category are largely independent of reconstruction method. The only discrepancy between the four reconstructions was one extra false positive result for the 'Sharp 1' reconstruction (the relatively large resultant 16.6% drop in specificity is caused by the low number of patients in this category). It is therefore reasonable to state that the overall diagnosis results for the smallest weight category were independent of the reconstruction method.

The '25 < BMI ≤ 35' weight category results follow a similar pattern as those noted for Observer 1: all 5 metrics largely worsen as the reconstruction sharpness increases, with 'Sharp 3' being an exception for all but sensitivity.

Whilst the results for the 'BMI > 35' category are of limited value, it is of interest that 'GE Clin' produced the greatest sensitivity and diagnostic accuracy results.

11.4.3 Lesion Detection Results

Table 11.9 summarises each observers' lesion detection performance for each reconstruction method. It should be noted that this table does not account for observer confidence in individual lesion marks.

		'GE Clin'	'Sharp 1'	'Sharp 2'	'Sharp 3'
Observer 1	True	31/45 (68.9%)	32/45 (71.1%)	29/45 (64.4%)	23/45 (51.1%)
	False	9	14	15	4
Observer 2	True	27/45 (60.0%)	29/45 (64.4%)	29/45 (64.4%)	28/45 (62.2%)
	False	0	3	3	0
Total	True	58/90 (64.4%)	61/90 (67.8%)	58/90 (64.4%)	51/90 (56.7%)
	False	9	17	18	4

Table 11.9: Summary of lesion detection performance, per observer

The 'Total' results in Table 11.9 demonstrate that 'Sharp 1' produced the greatest proportion of correctly identified lesions (67.8%), while 'Sharp 3' produced the least (56.7%). However, 'Sharp 1' produced 17 false positive lesions (second only to 'Sharp 2'), while 'Sharp 3' produced the lowest number of false positive lesions.

Observer 1's lesion detection performance varied with reconstruction sharpness. 'Sharp 1' produced the greatest number of correctly identified lesions, but this was at the expense of the second largest number of false positive lesions. 'Sharp 3' produced the lowest number of both correctly identified lesions and false positive lesions.

In contrast to Observer 1, Observer 2 correctly identified similar numbers of lesions for all four reconstruction methods; however, it is of interest that fewer lesions were seen on 'GE Clin' than any of the sharper reconstructions. 'Sharp 1' and 'Sharp 2' produced the greatest number of correctly identified lesions, but also produced the greatest number of false positive lesions (no false positives were seen on either 'GE Clin' or 'Sharp 3').
Figure 11.6 plots separate FROC curves for both observers.

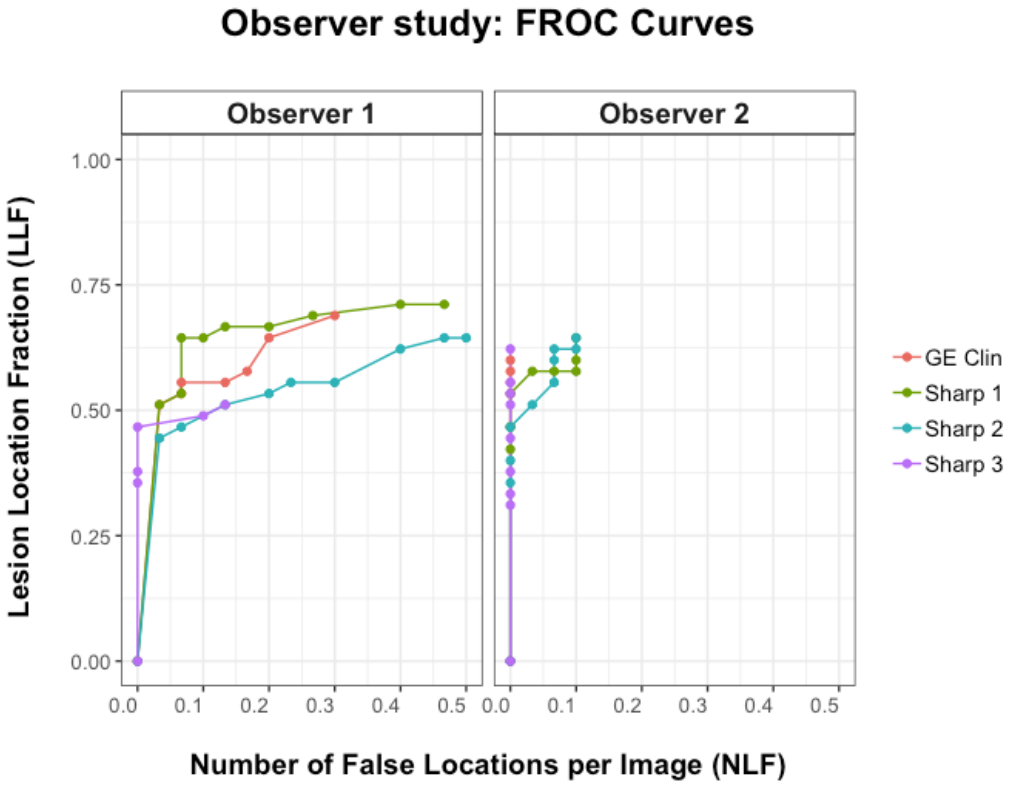


Figure 11.6: FROC curves for lesion detection task, per observer

Observer 1's curves for all four reconstructions reach a greater point on the x-axis than Observer 2's curves as a result of the greater number of false positives reported by Observer 1. Observer 1's results for 'Sharp 1' reach a greater point on the y-axis than any other result set, reflecting the superior fraction of true lesions found. The increased number of data points on Observer 1's curves for 'Sharp 1' and 'Sharp 2' indicate that Observer 1 was more willing to report lesions with a wider range of confidence scores when these reconstructions were used.

Observer 2's results for all four reconstructions appear to be reasonably similar, reflecting the results in Table 11.9. The relatively small number of data points in each of Observer 2's curves demonstrates Observer 2 did not report lesions over as wide a range of confidence levels as Observer 1.

Figure 11.7 plots separate wAFROC curves for both observers, whilst Table 11.10 compares the wAFROC AUC for each reconstruction.

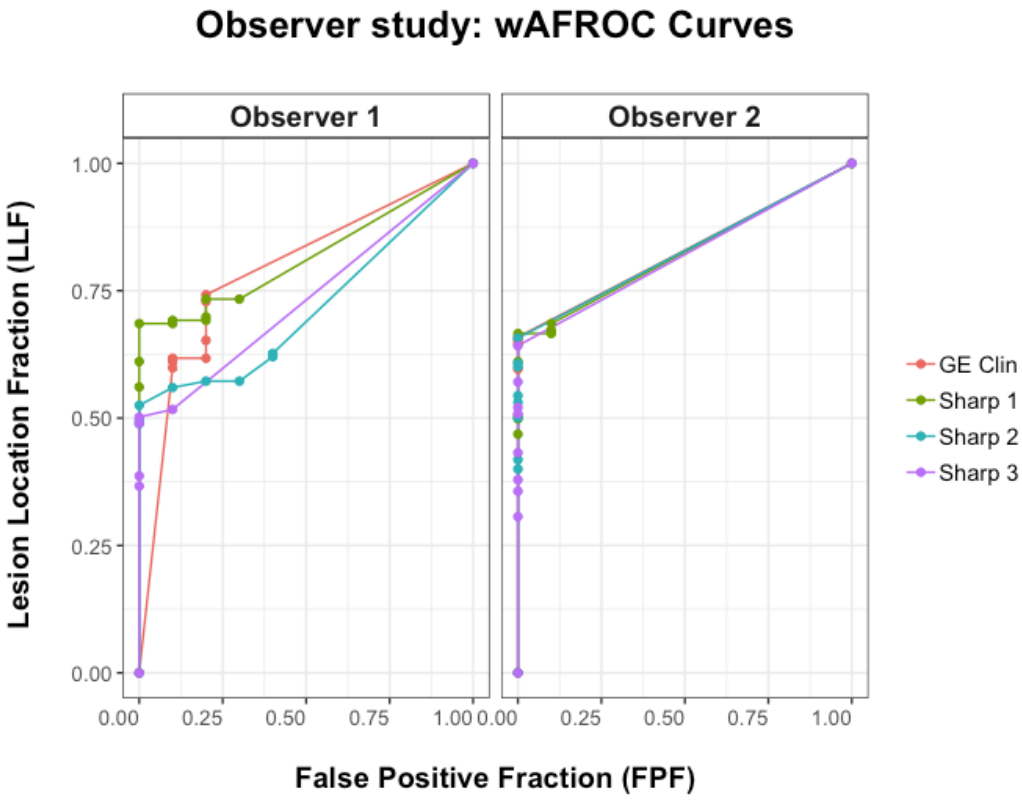


Figure 11.7: wAFROC curves for lesion detection task, per observer

		'GE Clin'	'Sharp 1'	'Sharp 2'	'Sharp 3'
AUC	**Observer 1**	0.788	0.818	0.716	0.734
	Observer 2	0.830	0.825	0.829	0.821
	Average	0.809	0.822	0.773	0.778

Table 11.10: AUC figures of merit from wAFROC curves

Observer 1's wAFROC curves appear to show that 'Sharp 1' is the superior reconstruction method, with 'GE Clin' in second place. This is confirmed by the AUC results: 'Sharp 1' has the greatest AUC (0.818). 'Sharp 2' is the inferior reconstruction, with AUC = 0.716.

Observer 2's wAFROC curves for all four reconstructions appear to be very similar. This is confirmed by the AUC results (all four AUCs are within 0.01 of each other). 'GE Clin' has the largest AUC by a margin of 0.001, with 'Sharp 2' in second place.

AUCs from the combined observer data sets confirm that 'Sharp 1' had the largest AUC, followed by 'GE Clin', 'Sharp 3' and finally 'Sharp 2'. However, statistical significance testing on the wAFROC data demonstrated there are no significant differences between the reconstruction methods. This remains the case when assessing each observer individually.

11.5 Discussion

The ‘GE Clin’ reconstruction was the preferred reconstruction in terms of both image quality and reporting confidence scores for both observers. Both image quality and confidence scores generally decreased as image sharpness increased. This was an expected result: the resultant increase in image noise reduced the observers’ perception of image quality, and therefore decreased their confidence in using the image for diagnosis. Furthermore, both observers had only used the ‘GE Clin’ reconstruction in clinical reporting prior to their participation in this study and had no experience of any of the sharper reconstructions. This may have introduced some preferential bias into their qualitative image quality assessments. It is of interest, however, that there were no statistically significant differences in confidence scores between ‘GE Clin’ and ‘Sharp 1’ for either observer. These results suggest that ‘Sharp 1’ could be used instead of ‘GE Clin’ without significantly affecting the observers’ perception of image quality or their confidence in using the reconstruction for liver diagnosis.

Patient BMI was shown to influence the relative performance of each reconstruction in terms of both image quality and reporting confidence. The ‘BMI $\leq$ 25’ category demonstrated no significant differences in image quality between ‘GE Clin’ and either ‘Sharp1’ or ‘Sharp 2’. Furthermore, the confidence results for the BMI $\leq$ 25 appeared to be unaffected by the choice of reconstruction; there were no significant differences between any of the reconstructions. These results suggest that for the smallest patients (BMI $\leq$ 25), ‘Sharp 1’ or ‘Sharp 2’ could be used instead of ‘GE Clin’ without significantly affecting the observers’ image perception. In contrast, the ‘25 < BMI $\leq$ 35’ category demonstrated a clear reduction in confidence as a result of increasing reconstruction sharpness. It is of interest, however, that ‘25 < BMI $\leq$ 35’ demonstrated no significant differences between ‘GE Clin’ and ‘Sharp 1’ in either image quality or reporting confidence. This suggests that ‘Sharp 1’ could be used instead of ‘GE Clin’ for patients with BMI $\leq$ 35 without significantly affecting the observers’ image perception.

The overall diagnosis results (normal versus abnormal liver) for each observer are summarised in Table 11.7. Given that Observer 1’s image quality and reporting confidence decreased as reconstruction sharpness increased, one would expect a corresponding degradation in diagnosis metrics. Whilst this was largely the case for Observer 1, there was a paradoxical increase in the ‘Sharp 3’ diagnosis metrics for all but sensitivity. Furthermore, ‘Sharp 3’ produced greater specificity and NPV than all other reconstructions. One would expect that increased image noise may increase the risk of false positives, which would in turn cause a decrease in both specificity and NPV. A likely reason for the paradox in

Observer 1's results is that they found the 'Sharp 3' reconstructions to be of such low quality, they became less likely to attempt to identify potential low-confidence lesions, thus reducing the false positive rate (hence increasing specificity and NPV) but also reducing the true positive rate (hence reducing sensitivity).

Table 11.7 demonstrated very little difference in Observer 2's diagnostic performance between the four reconstructions: a single false positive in a 'Sharp 2' reconstruction caused the only discrepancies. This was consistent with Observer 2's confidence results, which were very similar for all four reconstruction methods (no statistically significant differences).

Diagnosis results were also analysed for different BMI categories. When considering all diagnosis results together (Tables 11.7 and 11.8), one can infer the following:

- Observer 2's overall diagnosis results were independent of reconstruction method (regardless of patient BMI)
- Observer 1's results for patients with BMI $\leq$ 25 were independent of reconstruction method
- Observer 1's results for patients with BMI > 25 generally worsened with increasing reconstruction sharpness, with the exception of the paradoxical 'Sharp 3' results.

The overall diagnosis results (normal versus abnormal) did not provide any evidence to justify the use of a sharper reconstruction then 'GE Clin'. In particular, 'GE Clin' appeared to produce the best results for patients with BMI > 25 (allowing for the paradoxical 'Sharp 3' results). However, this type of analysis does not consider decision-making at the level of individual lesions. Furthermore, it does not take into consideration the confidence level of individual false positive lesions that resulted in overall false positive diagnoses; a clinician applying a lower confidence threshold in order to maximise lesion detection would risk an increased level of overall diagnosis false positives in this particular analysis.

The final part of the analysis assessed lesion detection. This analysis was not broken down into the different BMI categories due to the small number of patients in each category. Both observers identified almost the same number of true lesions but had different numbers of false positives: Observer 1 identified 42 false positives, while Observer 2 identified only 6 false positives. This suggests the observers had different approaches to identifying lesions. The confidence scores assigned to false positive lesions were interrogated and demonstrated Observer 1 was more likely to identify lesions with low confidence than Observer 2. For example, Observer 1 identified 18 false positive lesions with confidence levels of 50% or below (median false positive confidence was 60%) while Observer 2 identified only one false positive with 50% confidence (median false positive confidence

was 85%). This explains why Observer 1 had lower specificity results than Observer 2 in the previous section.

Table 11.9 summarises the true and false positive lesions identified by each observer, for each reconstruction (lesion confidence was not accounted for in this table). Taking the previous reporting confidence results into account, one would expect Observer 1's results to have some dependence upon reconstruction, with Observer 2's results being more consistent across all reconstructions. This is largely the case for the true positives: Observer 2's results were reasonably consistent (maximum discrepancy of 2 lesions), while Observer 1 typically detected fewer true lesions using the sharper reconstructions (maximum discrepancy of 9 lesions). Interestingly, while both observers rated 'GE Clin' highest in terms of image quality and reporting confidence, this reconstruction did not provide the optimum true positive detection rate for either observer:

- Observer 1 detected most true positive lesions using 'Sharp 1', with 'GE Clin' in second place
- Observer 2 detected most true positive lesions using both 'Sharp 1' and 'Sharp 2', with 'GE Clin' producing the lowest detection rate.

As stated previously in the discussion, one may have expected the false positive detection rate to increase as reconstruction sharpness, and hence image noise, increased. This was largely the case for both observers, with the exception of 'Sharp 3'. The effects of BMI on image quality were demonstrated by the patient images in Figure 11.3. Noise within the largest patient's 'Sharp 3' reconstruction makes it difficult to identify the relatively large confirmed lesion which is more easily visualised on the 'GE Clin' reconstruction. The same may be true to a lesser degree for the medium sized patient's liver lesion. It is likely that both observers found 'Sharp 3' to be of such low quality for the larger patients, they were unable to identify the same false lesions due to the increased noise levels. Furthermore, the increased noise in 'Sharp 3' disproportionately affected Observer 1's true positive rate compared to Observer 2 (Observer 1 had lower confidence in 'Sharp 3' compared to Observer 2).

Although the FROC curves produced a useful visual representation of each observers' lesion detection performance (largely confirming the results in Table 11.9 whilst partially taking lesion confidence into account), the wAFROC curves and AUCs produced a more complete analysis of the available data (true negatives were taken into account, as well as individual lesion confidence and true lesion weightings). The wAFROC AUCs for the complete data set confirmed that 'Sharp 1' had the superior lesion detection performance of the four reconstructions; however, there were no statistically significant differences

between the four reconstruction methods (even when each observer was assessed individually).

When deciding which reconstruction is optimal, one must take into consideration the specific aims of the clinical task. For example, is maximising the detection of true positive lesions more important than minimising the number of false positive lesions detected? In liver imaging, a missed lesion may result in a patient missing out on appropriate treatment at an earlier stage in their disease or potentially cause them to undergo painful and expensive curative treatment which would ultimately prove to be futile. On the other hand, a false positive lesion in the liver could cause a patient to be denied potential curative surgery or to undergo unnecessary surgical or ablative procedures.

Colorectal cancer patients typically undergo MRI scans as well as PET scans. Other oncology patients whose PET scans reveal previously unknown liver metastases are highly likely to be scheduled for an MRI as a result of the PET findings, prior to any change in clinical management. It is therefore unlikely a false positive liver lesion on a PET scan would directly lead to an inappropriate change in patient management. One may therefore justify the increased risk of false positives if the number of true positives can be improved. With this in mind, the results of this observer study suggest that the 'Sharp 1' reconstruction may be more appropriate than the currently used 'GE Clin' for lesion detection.

Both observers preferred the 'GE Clin' reconstruction in terms of image quality and reporting confidence despite detecting more true lesions using 'Sharp 1'. Whilst the differences in reporting confidence between these two reconstructions were not statistically significant, it may be useful to present both reconstructions to reporting clinicians for a period of time to build up their confidence in the sharper reconstruction. Viewing both reconstructions together may even prove to be a useful long-term solution if clinicians find the combination to be beneficial.

Phantom studies in Chapters 9 and 10 suggested that 'Sharp 2' would maximise detectability of small liver lesions within smaller patients. This chapter found that the use of 'Sharp 2' would have no significant effect upon perceived image quality or reporting confidence for patients with BMI < 25. Whilst the lesion detection study appeared to favour 'Sharp 1' (albeit with no statistical significance), the patient data set was dominated by larger patients (70% of the patients had BMI > 25). It is therefore possible that smaller patients may benefit from a further increase in image sharpness than that offered by 'Sharp 1'.

Studies by Schaefferkoetter et al [8], Morey et al [157] and Kadrmas et al [220] assessed general oncology lesion detection using Siemens Biograph PET-CT systems and

mathematical observers. All three studies used slightly different combinations of post-reconstruction filtering and effective iterations to maximise lesion detection using TOF+PSF algorithms; respectively, the combinations were as follows: 24 effective iterations with no filtering, 90 effective iterations with 3mm filter (2mm voxels) and 84 effective iterations with 2.4mm filter (2mm voxels). None of these studies specifically targeted liver lesion detection or assessed the GEMS Discovery PET-CT system, and so are not directly comparable to the results of this study.

A 2017 study by Adler et al [223] evaluated a novel phantom developed for assessing minimum lesion detectability. Several different PET-CT systems were used in the study, including the GEMS Discovery 710 (similar to the Discovery 690 in terms of OSEM performance). The authors state they used the manufacturers' recommended reconstruction settings for each system, which were as follows for the GEMS Discovery 710: TOF+PSF, 120 effective iterations, 2mm transaxial filter (no z-axis filter or voxel size specified). It is of interest that these settings are more similar to the 'Sharp 1' reconstruction (108 effective iterations, no filtering) recommended by this chapter than the 'GE Clin' reconstruction (54 effective iterations, 4mm filter).

The main limitation of this study was the absence of a true gold standard. Best endeavours were made with the available correlative imaging (using MRI in particular, as this is the current imaging gold standard for liver lesions), pathology and follow-up data to determine the absolute truth for each patient used in this study. This methodology has been used by several studies [243], [246], [247]. However, unlike phantom data, the absolute truth within a patient study can never be fully known. It is even possible, if unlikely, that some of the false positive lesions identified may in fact have represented true lesions at the time of the PET scan.

Another limitation of this study was the relatively small number of patients: the potential for demonstrating a statistically significant difference between the four reconstructions, in terms of lesion detection performance, would be improved by undertaking a study with a larger sample size. This study demonstrated the effects of patient BMI upon relative reconstruction performance; a larger study should therefore include larger numbers of patients in all BMI categories to provide more meaningful analysis (e.g. wAFROC analysis could be performed on each BMI category individually). Based on the initial findings presented in this chapter, it is likely such a study would indicate different reconstructions are optimal for different BMI categories, with smaller patients being potentially more amenable to sharper reconstructions.

A further limitation in this study was the small number of observers, and their bias. It is of note that there appeared to be a bigger difference between the two observers than the reconstruction algorithms: this may be a result of the inclusion of low confidence lesions. Furthermore, this study did not perform any analysis of intra-observer reliability. Any future studies should incorporate more observers and assess each observers' reliability for each reconstruction method.

Patient data used in this study also demonstrated a limited number of small, challenging lesions. Ideally, a larger patient study would include a larger number of lesions that, though detectable, would provide a challenge to the observers in order to better differentiate the reconstructions in terms of lesion detection performance. Future analysis could also directly relate lesion detection performance to lesion size measurements (e.g. measured using MRI data), lesion SUV measurements and background liver noise measurements (e.g. image roughness).

11.6 Conclusions

The patient observer study has demonstrated liver lesion detection may be improved by using the 'Sharp 1' reconstruction instead of the reconstruction parameters suggested by GEMS (the 'GE Clin' reconstruction'), which are currently in clinical use. This would not result in a significant change in reporting confidence, although there may be a perceptible reduction in image quality for patients with BMI > 25. It may therefore be useful to present both reconstructions to reporting clinicians for a period of time to build up reporting confidence in the sharper reconstruction. This chapter therefore concludes that both 'GE Clin' and 'Sharp 1' reconstructions should be used together when interpreting ^{18}F-FDG liver images, to improve small lesion detection.

Although the liver lesion detection analysis in this chapter did not directly assess the effects of patient BMI, this chapter has demonstrated it is reasonable to suggest that PET reconstruction could be further optimised by tailoring the reconstruction to patient BMI. A larger patient study involving larger numbers of patients in each relevant BMI category would be required to optimise liver lesion detection for each BMI category.

Whilst this chapter has assessed lesion detection within the liver, the observer study methodology used in this chapter can be followed for different clinical applications with multifocal truth; for example, optimising lesion detection in lung or head and neck imaging studies. PET reconstructions may therefore be tailored to different clinical tasks and different patient groups (e.g. BMI categories).

All patients included in this study followed the same ^{18}F-FDG injection protocol, regardless of patient weight. The resultant inconsistency in patient activity concentrations therefore contributed to the relative noise increase as patient BMI increased. As many imaging centres now use weight-based injection protocols [14], [186], future lesion detection studies should account for this. This may affect the potential for increasing the sharpness for smaller patients, as a reduction in activity concentrations for these patients will lead to increased image noise. Furthermore, all patients were reconstructed using the same transaxial FOV, regardless of size. Further work should investigate the effects of reducing the FOV for smaller patients, thus reducing voxel sizes. The resultant increase in both spatial resolution and image noise would likely influence the optimum combinations of reconstruction parameters for specific clinical tasks.

Chapter 12 : Conclusions

The primary aim of this thesis was to develop a generic methodology to assess and optimise PET image acquisition and reconstruction that can be applied to any clinical application. Although a GEMS Discovery 690 PET-CT system was used for all practical work in this thesis, the generic methodology was intended to be useful for all PET-CT systems. The secondary aim was to optimise image reconstruction using the GEMS Discovery 690 PET-CT system, specifically for the detection of small liver lesions in ^{18}F-FDG-PET oncology imaging.

12.1 Implications for the Development of a Generic Methodology for Assessment and Optimisation of Clinical PET Image Reconstruction

To the author's knowledge, there is no established methodology for the assessment and optimisation of PET system performance in a clinically relevant manner. Such a methodology would be particularly useful when new technological advancements become available. For example, the use of PSF remains controversial in clinical imaging: some studies have concluded no post-reconstruction filtering is required when PSF is applied, whilst others recommended the use of filters to mitigate potential Gibbs artefacts. Some studies have recommended PSF should not be used at all, particularly for quantitative studies. This thesis has assessed the effects of various reconstruction parameters upon noise, spatial resolution and lesion detection using clinically relevant phantom techniques, with a particular focus upon liver imaging; e.g. the activity concentrations used in the spatial resolution and body phantoms were chosen to be representative of clinical liver imaging

with 400MBq ^{18}F-FDG. A patient observer study was then undertaken to clinically assess the recommendations made based on the phantom data.

Although the primary focus of this thesis was image reconstruction, there were acquisition issues that also merited attention. GEMS recommend the use of 23% slice overlaps between acquisition frames; however, this had not been justified by independent peer-review studies. Chapter 4 compared image quality obtained using a 23% overlap with that of the maximum 49% overlap. This chapter concluded there was minimal detectable influence on clinical image quality in terms of noise, mean activity concentrations or lesion contrast recovery. All subsequent imaging work undertaken throughout this thesis therefore used a 23% overlap. Furthermore, this chapter proposed an amendment to the EANM guidelines to account for the relationship between image noise and slice overlap when calculating minimum patient ^{18}F-FDG injection activities, which has been published in a peer-reviewed journal [179]. The GEMS suggested overlap recommendations were therefore found to be satisfactory for lesion detection under current clinical conditions (400MBq ^{18}F-FDG imaging with four-minute acquisition frames). Slice overlap would, however, require reassessment for different clinical imaging scenarios.

Image noise was assessed using two phantoms: 20cm diameter ^{68}Ge uniform phantom (Chapter 6) and the NEMA body phantom containing simulated lesions (Chapter 9). Whilst both phantoms demonstrated the effects of altering reconstruction parameters, the uniform phantom was too small to demonstrate the advantages of TOF in whole-body liver imaging (although results were relevant for e.g. head and neck or paediatric imaging). Furthermore, the use of a uniform phantom provides an over-simplistic imaging scenario that is not representative of clinical imaging. Phantoms containing simulated lesions, like the NEMA body phantom, provide a more complex imaging scenario closer to that of clinical imaging. This thesis therefore recommends that image noise should be assessed using the background chambers of phantoms containing simulated lesions, and background chambers should be size-matched to the clinical task in question. Furthermore, this thesis recommends assessing both voxel-to-voxel and region-to-region noise where possible.

No technique is established for the measurement of clinical spatial resolution, yet this is critical to image optimisation. Spatial resolution is challenging to assess in a clinically relevant manner, particularly when iterative reconstruction methods are used. Chapter 7 developed a methodology for measuring spatial resolution in a clinically relevant manner. This chapter concluded by recommending the use of line sources within a background activity. Furthermore, this chapter demonstrated complex corrections for the background activity were not required due to the inherent limitations of PET spatial resolution; a relatively simple method of background subtraction was sufficient. Additionally, the

assessment technique allowed reliable FWHM measurement using the full transaxial FOV when a 256 matrix was used; results were therefore relevant to clinical liver imaging.

This thesis has also shown that vendor implementations of reconstruction parameters may not always function as intended. An unexpected relationship (unknown to even senior GEMS engineers) was discovered between matrix size, Gaussian filter width and image noise: at filter widths between approximately 3mm and 9mm, smaller voxels produced less image noise than larger voxels. Communications with GEMS established this was caused by their Gaussian filter implementation. This may be of clinical relevance, as GEMS suggest using a 4mm Gaussian filter for clinical reconstructions. This thesis therefore recommends that the effects of reconstruction parameters should not be assessed in isolation; in particular, the effects of Gaussian filtering and voxel sizes on image quality should be assessed together.

Chapter 9 assessed the effects of reconstruction parameters upon lesion detection using the widely available NEMA body phantom. This phantom was shown to be a reasonable representation of smaller patients; however, larger background chambers should be used to represent larger patients. A novel SNR methodology, based upon region-to-region noise instead of the widely-used voxel-to-voxel noise, was also assessed and was shown to be a better predictor of observer preference than the traditional SNR method. However, the chapter concluded that dual metric $HCRC_{mean}$ and noise analysis provided the best predictions of human observer preferences for small lesion detection. These methods of quantitative assessments should be included when optimising PET reconstruction techniques for other clinical applications

Whilst this thesis has assessed lesion detection within the liver, the observer study methodology used in Chapter 11 can be followed for different clinical applications with multifocal truths; for example, optimising lesion detection in lung or head and neck imaging studies. PET reconstructions may therefore be tailored to different clinical tasks and different patient groups (e.g. BMI categories). This thesis therefore recommends the use of FROC and wAFROC analysis to compare lesion detection performance of different reconstructions; such analyses allow observers to identify an unrestricted number of lesions for each patient and is therefore more clinically relevant than conventional LROC studies. However, a larger number of patients than used in this thesis would be required to demonstrate statistically significant differences between different reconstruction methods.

12.1.1 Recommended Methodology for Assessment and Optimisation of Clinical PET Reconstruction

This thesis therefore concludes the following evidence-based approach should be taken to optimising PET clinical reconstruction:

Step1: Slice Overlap (if user selectable option on PET system model):

- A complex phantom simulating lesions within a background activity (for example, the NEMA IEC body phantom) should be filled with activity concentrations relevant to both the clinical task and local injection activity protocol (total phantom activity should be chosen to be conservative). Spheres should be used to represent low contrast lesions, e.g. 2:1 ratio. The phantom should be acquired with spheres in the overlap area. The following overlap sizes should be assessed (at a minimum):
    - Maximum possible overlap (likely to be approximately 50%)
    - 23% overlap (shown to be the minimum acceptable in this thesis)
    - Manufacturer recommended overlap (if different from those above)
- Phantoms should be reconstructed using manufacturer's suggested parameters
- Qualitative assessment should compare sphere detectability for different overlap settings. This may be sufficient to determine if the lower overlap settings are acceptable.
- Quantitative assessment of image quality should follow the analysis procedure detailed in Step 3. This may be used as evidence to show the chosen overlap setting meets international guidance requirements for image noise and lesion recovery.

Step 2: Spatial resolution:

- A line source phantom within a background activity should be used, similar to the NEMA 1994 line source phantom (phantom should ideally have at least two line sources). The background chamber should be filled with an activity concentration relevant to the clinical task. Line sources should be filled with an approximate 200:1 line-to-background ratio. The phantom should be positioned such that one point source is central and one point source is peripheral, to provide an indication of resolution uniformity throughout the FOV.
- Phantom data should be reconstructed using the maximum transaxial FOV relevant to the clinical task. Transaxial voxel sizes of 2.73mm or less should provide adequate sampling for reliable FWHM measurement.
- Reconstructed images should be analysed by drawing line profiles through the line source images and extracting the voxel values. A background ROI should

be used to extract the mean background voxel value. Line profile voxel values should then be corrected for background activity by subtracting the mean background voxel value.

- Corrected line profiles should be fitted to Gaussian functions, and the FWHM of the Gaussian function should then be calculated.
- FWHM at both FOV positions should be compared as reconstruction parameters vary; differences in FWHM at both FOV positions should be minimised.

Step 3: Image Noise and Lesion Detection (and/or Quantitative Accuracy):

- A complex phantom simulating lesions within a background activity (for example the NEMA IEC body phantom) should be filled with activity concentrations relevant to the clinical task and local injection activity protocol (total phantom activity should be chosen to be conservative). The phantom should be acquired with spheres in the overlap area, using the overlap size determined in Step 1.
- Phantom data should be reconstructed using a range of parameter combinations. The total number of reconstructions to be assessed may be narrowed down based on spatial resolution results, e.g. using only parameter combinations that produced uniform spatial resolution at the different FOV positions.
- Qualitative assessment should compare image quality of the overlap slices containing simulated lesions: this may be sufficient to rule out further parameter combinations if the clinical task is lesion detection (for example high noise reconstructions when high numbers of iterations are applied, or low resolution reconstructions when greater filtering is applied).
- The following image metrics should be analysed to further aid selection of reconstruction parameters and to provide quantitative evidence justifying the final reconstruction choice(s):
    - Coefficient of Variation: voxel-to-voxel noise, using large ROIs (allows comparison with EANM/EARL noise guidelines)
    - Background Variation: region-to-region noise, using ROIs matched to each of the lesion spheres used in the phantom
    - HCRCs of spheres (allows comparison with EANM/EARL guidelines; provides a measure of quantitative accuracy of lesions)
    - SNR of spheres, using sphere-matched Background Variation results as the noise metric
- Dual metric HCRC versus noise (using both COV or Background Variation) and SNR should provide a reasonable prediction for human observer preference.

Step 4: Patient Observer Study

- Reconstruction strategies identified by the preceding steps as providing optimal image quality should ideally be assessed using patient data; parameters suggested by the manufacturers should be included for comparison. This methodology can be followed to assess patients with potentially multifocal disease and available gold standard data.
    - Observers should record scores for image quality and reporting confidence
    - Observers should report the patient images (or organ of interest) as being normal or abnormal)
    - Observers should identify all lesions they observe, recording their positions and stating their confidence in each individual lesion
- FROC and wAFROC analysis should be undertaken to establish which reconstruction method provided the optimal lesion detection performance. This should be compared with the observer image quality and confidence results.
- It may be appropriate to use two reconstruction parameters when reporting clinical data; one reconstruction to aid general image interpretation and one to optimise the particular task being assessed (e.g. lesion detection or quantitative accuracy).

12.1.2 Further work

As PET hardware and software both continue to evolve (e.g. digital PET detectors, larger axial imaging FOVs and the development of more sophisticated reconstruction techniques), the small lesion detection performance of PET-CT imaging systems will improve. The smallest sphere in the current NEMA body phantom has an inner diameter of 9.9mm; however, a smaller set of spheres compatible with the NEMA phantom are available commercially, with inner diameters as small as 3.95mm. The use of these sub-centimetre spheres would provide a more challenging lesion detection task. Furthermore, the introduction of unpredictable sphere positioning within the phantom would further improve the phantom's relevance to clinical imaging. Work is therefore ongoing to produce a novel lesion detection phantom where the sizes, numbers and positions of spheres can be altered, and remain unknown to human observers assessing the images. The use of different background chambers will also be investigated, to better represent whole-body imaging of patients with varying BMI.

The spatial resolution phantom could also be improved. The use of more lines sources at different positions within a background chamber will also be investigated. Furthermore, the use of the ^{22}Na point source will be revisited; this may provide a more reproducible

methodology when the point source activity concentration has decayed to a more clinically relevant level.

12.2 Implications for Liver Lesion Detection on the GEMS Discovery 690 PET-CT System

There is no established optimum reconstruction method for detecting small lesions within the liver in ^{18}F-FDG PET imaging. The secondary aim of this thesis was therefore to recommend an optimised reconstruction strategy for small liver lesion detection. Phantom experiments were designed to be representative of liver imaging. Patient liver data were then used to clinically assess lesion detection performance of four different reconstruction strategies, chosen as a result of the phantom work.

Spatial resolution analysis demonstrated that, contrary to recent published data using non-clinical measurement techniques (e.g. using sources in air instead of background activity), PSF was shown to degrade resolution at lower iterations, and required over 180 iterations to improve spatial resolution. TOF-only produced superior spatial resolution when up to 90 iterations were applied. Applying PSF and TOF together produced superior spatial resolution when between 90 and 180 iterations were applied. Lesion detection and noise analysis then demonstrated that applying both PSF and TOF together with 108 iterations and no filtering appeared to optimise detection of the smallest sphere. A comparison between the phantom's dimensions and those of patients with a range of body habitus demonstrated that the NEMA phantom was a reasonable approximation for liver imaging in patients with BMI < 35; larger patients would be better represented by a phantom with a larger background chamber.

An observer study was conducted to assess three different reconstruction strategies selected using the phantom data, together with the GEMS suggested reconstruction strategy. Reassuringly, the parameters suggested by GEMS were shown to be close to optimal for liver lesion detection. However, while both observers preferred the GEMS suggested reconstruction in terms of reporting confidence and perceived image quality, results demonstrated a sharper reconstruction ('Sharp 1') may improve the detection of liver lesions. This chapter concluded that to improve lesion detection and maintain reporting confidence, both the GEMS suggested reconstruction and 'Sharp 1' reconstructions should be used together when interpreting ^{18}F-FDG liver images.

12.2.1 Recommended Reconstruction Strategy for ^{18}F-FDG Liver Imaging

This thesis therefore recommends the following reconstruction strategies are used together for all patients undergoing 400MBq ^{18}F-FDG whole-body imaging on the GEMS Discovery 690 PET-CT system:

- The GEMS suggested reconstruction parameters, for general image interpretation:
    - PSF+TOF, 54 effective iterations, 4mm Gaussian filter, Standard z-axis filter, 192 matrix with 700mm FOV (3.65mm voxel size)

- The 'Sharp 1' reconstruction parameters, for optimal liver lesion detection:
    - PSF+TOF, 54 effective iterations, no filtering, 256 matrix with 700mm FOV (2.73mm voxel size)

12.2.2 Further work

Results in this thesis suggest that, under the current injected activity protocol, lesion detection in smaller patients (BMI < 35) may benefit from even sharper reconstructions. Reconstruction protocols could therefore be optimised further by tailoring to patient BMI.

Weight-based ^{18}F-FDG injection protocols are now commonly used. Future studies should therefore account for reduced activity concentrations in smaller patients. This may affect the potential for increasing reconstruction sharpness for smaller patients, as a reduction in activity concentrations for these patients will lead to increased image noise.

Reconstructions could also be improved by tailoring the transaxial FOV to patient size. The resultant increase in both spatial resolution and image noise would likely influence the optimum combinations of reconstruction parameters for specific clinical tasks.

The work performed in this thesis has demonstrated that there is no single method for optimum image acquisition and reconstruction across the board; many external factors influence image quality, making it challenging to decide how to approach patient imaging. It is therefore hoped that the findings presented here will assist those creating PET acquisition and reconstruction protocols to accomplish what can appear to be be a daunting task.

References

[1] D. Hanahan and R. A. Weinberg, “Hallmarks of Cancer: The Next Generation,” *Cell*, vol. 144, no. 4, pp. 646–674, 2011.

[2] F. H. Martini, *Fundamentals of Anatomy and Physiology*, 7th Editio. Pearson Education Inc., 2006.

[3] L. A. Torre, F. Bray, R. L. Siegel, and J. Ferlay, “Global Cancer Statistics, 2012,” vol. 65, no. 2, pp. 87–108, 2015.

[4] J. Ferlay, M. Colombet, I. Soerjomataram, C. Mathers, D. M. Parkin, M. Piñeros, A. Znaor, and F. Bray, “Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods,” *Int. J. Cancer*, 2018.

[5] M. Naghavi, A. A. Abajobir, R. W. Kimokoti, Y. Kinfu, N. Kissoon, M. Kivimaki, A. K. Knudsen, J. A. Kopec, S. Kosen, P. A. Koul, A. Koyanagi, B. K. Defo, X. R. Kulikof, G. A. Kumar, P. Kumar, M. Kutz, H. H. Kyu, D. K. Lal, R. Lalloo, T. L. N. Lambert, Q. Lan, V. C. Lansingh, A. Larsson, P. H. Lee, J. Leigh, J. Leung, M. Levi, Y. Li, D. L. Kappe, X. Liang, M. L. Liben, S. S. Lim, A. Liu, P. Y. Liu, Y. Liu, R. Lodha, G. Logroscino, S. Lorkowski, P. A. Lotufo, R. Lozano, T. C. D. Lucas, S. Ma, E. R. K. Macarayan, E. R. Maddison, M. M. Abd El Razek, M. Majdan, R. Majdzadeh, A. Majeed, R. Malekzadeh, R. Malhotra, D. C. Malta, H. Manguerra, T. Manyazewal, C. C. Mapoma, L. B. Marczak, D. Markos, J. Martinez-Raga, M. Tavakkoli, N. Taveira, A. D. Lopez, and C. J. L. Murray, “Global, regional, and national age-sex specifc mortality for 264 causes of death, 1980-2016: A systematic analysis for the Global Burden of Disease Study 2016,” *Lancet*, vol. 390, no. 10100, pp. 1151–1210, 2017.

[6] R. L. Wahl, H. Jacene, Y. Kasamon, and M. a Lodge, “From RECIST to PERCIST: Evolving Considerations for PET response criteria in solid tumors.,” *J. Nucl. Med.*, vol. 50 Suppl 1, no. 5, p. 122S–50S, May 2009.

[7] World Health Organization, “Guide to Cancer: Early Diagnosis,” 2017.

[8] J. Schaefferkoetter, M. Casey, D. Townsend, and G. El Fakhri, “Clinical impact of time-of-flight and point response modeling in PET reconstructions : a lesion detection study,” *Phys Med Biol*, no. 58, pp. 1465–1478, 2013.

[9] J. Brierley, M. Gospodarowicz, and B. O’Sulivan, “The principles of cancer staging,” *Ecancermedicalscience*, vol. 10, pp. 3–6, 2016.

[10] M. E. Phelps, “Positron emission tomography provides molecular imaging of biological processes.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, no. 16, pp. 9226–33, Aug. 2000.

[11] S. I. Ziegler, “Positron Emission Tomography: Principles, Technology, and Recent Developments,” *Nucl. Phys. A*, vol. 752, pp. 679–687, Apr. 2005.

[12] O. Warburg, “On the Origin of Cancer Cells,” *Science (80-.).*, vol. 123, no. 3191, pp. 309–14, 1956.

[13] R. J. DeBerardinis, J. J. Lum, G. Hatzivassiliou, and C. B. Thompson, “The Biology of Cancer: Metabolic Reprogramming Fuels Cell Growth and Proliferation,” *Cell Metab.*, vol. 7, no. 1, pp. 11–20, 2008.

[14] R. Boellaard, R. Delgado-Bolton, W. J. G. Oyen, F. Giammarile, K. Tatsch, W. Eschner, F. J. Verzijlbergen, S. F. Barrington, L. C. Pike, W. a. Weber, S. Stroobants, D. Delbeke, K. J. Donohoe, S. Holbrook, M. M. Graham, G. Testanera, O. S. Hoekstra, J. Zijlstra, E. Visser, C. J. Hoekstra, J. Pruim, A. Willemsen, B. Arends, J. Kotzerke, A. Bockisch, T. Beyer, A. Chiti, and B. J. Krause, “FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 42, no. 2, pp. 328–354, 2015.

[15] J. Czernin, M. Allen-Auerbach, D. Nathanson, and K. Herrmann, “PET/CT in Oncology: Current Status and Perspectives,” *Curr. Radiol. Rep.*, vol. 1, no. 3, pp. 177–190, 2013.

[16] IAEA, “IAEA Human Health Series No.9: Appropriate Use of FDG-PET for the Management of Cancer Patients,” 2010.

[17] IAEA, “IAEA Human Health Series No.27: PET / CT Atlas on Quality Control and Image Artefacts,” 2014.

[18] C. Lois, B. W. Jakoby, M. J. Long, K. F. Hubner, D. W. Barker, M. E. Casey, M. Conti, V. Y. Panin, D. J. Kadrmas, and D. W. Townsend, “An assessment of the impact of incorporating time-of-flight information into clinical PET/CT imaging.,” *J. Nucl. Med.*, vol. 51, no. 2, pp. 237–45, Feb. 2010.
[19] V. Bettinardi, L. Presotto, E. Rapisarda, M. Picchio, L. Gianolli, and M. C. Gilardi, “Physical performance of the new hybrid PET/CT Discovery-690.,” *Med. Phys.*, vol. 38, no. 10, pp. 5394–411, Oct. 2011.
[20] A. Rahmim and H. Zaidi, “PET versus SPECT: strengths, limitations and challenges.,” *Nucl. Med. Commun.*, vol. 29, no. 3, pp. 193–207, 2008.
[21] J. L. Herraiz, S. España, J. M. Udías, J. J. Vaquero, and M. Desco, “Statistical Reconstruction Methods in PET : Resolution Limit , Noise , Edge Artifacts and considerations for the design of better scanners,” *IEEE Nucl Sci Symp Conf Rec.*, pp. 1846–1850, 2005.
[22] A. Alessio and P. Kinahan, “PET Image Reconstruction,” in *Nuclear Medicine*, 2005, pp. 1–22.
[23] R. M. Manjeshwar, F. P. M. H. Jansen, and C. W. Stearns, “US 7381959 Technique for reconstructing PET scan images,” 2008.
[24] ARSAC, “Notes for Guidance on the Clinical Administration of Radiopharmaceuticals and Use of Sealed Radioactive Sources,” 2006.
[25] S. R. Cherry, T. Jones, J. S. Karp, J. Qi, W. W. Moses, and R. D. Badawi, “Total-Body PET: Maximizing Sensitivity to Create New Opportunities for Clinical Research and Patient Care,” *J. Nucl. Med.*, vol. 59, no. 1, pp. 3–12, 2017.
[26] S. I. Kabanikhin, “Definitions and examples of inverse and ill-posed problems,” *J. Inverse Ill-Posed Probl.*, vol. 16, no. 4, pp. 317–357, 2008.
[27] A. Brahme, *Comprehensive Biomedical Physics*. Newnes, 2014.
[28] M. Phelps, *PET: Physics, Instrumentation, and Scanners*. Springer, 2006.
[29] R. D. Badawi, “Aspects of Optimisation and Quantification in Three-Dimensional Positron Emission Tomography,” 1998.
[30] G. Tarantola, F. Zito, and P. Gerundini, “PET instrumentation and reconstruction algorithms in whole-body applications.,” *J. Nucl. Med.*, vol. 44, no. 5, pp. 756–69, May 2003.
[31] N. Zeraatkar, A. Rahmim, S. Sarkar, and M. R. Ay, “Development and Evaluation of Image Reconstruction Algorithms for a Novel Desktop SPECT System,” vol. 5, no. 2, pp. 120–133, 2017.
[32] H. Zaidi, *Quantitative Analysis in Nuclear Medicine Imaging*. Boston: Kluwer Academic Publishers, 2006.
[33] D. Wiant, J. A. Gersh, M. Bennett, and J. D. Bourland, “Evaluation of the spatial dependence of the point spread function in 2D PET image reconstruction using LOR-OSEM.,” *Med. Phys.*, vol. 37, no. 3, pp. 1169–1182, 2010.
[34] S. Stute, D. Benoit, A. Martineau, N. S. Rehfeld, and I. Buvat, “A method for accurate modelling of the crystal response function at a crystal sub-level applied to PET reconstruction,” *Phys. Med. Biol.*, vol. 56, no. 3, pp. 793–809, 2011.
[35] J. Prekeges, *Nuclear Medicine Instrumentation*. Jones and Bartlett, 2013.
[36] GE Healthcare, “VUE Point HD ™ - Bringing Accuracy to PET Reconstruction,” 2009.
[37] A. Rahmim, J. Qi, and V. Sossi, “Resolution modeling in PET imaging: Theory, practice, benefits, and pitfalls,” *Med. Phys.*, vol. 40, no. 6, p. 064301, 2013.
[38] F. E. Turkheimer, M. Veronese, and J. Dunn, *Experimental Design and Practical Data Analysis in Positron Emission Tomography*. 2014.
[39] J. E. Ortũno, G. Kontaxakis, J. L. Rubio, P. Guerra, and A. Santos, “Efficient methodologies for system matrix modelling in iterative image reconstruction for rotating high-resolution PET,” *Phys. Med. Biol.*, vol. 55, no. 7, pp. 1833–1861, 2010.
[40] H. M. Hudson and R. S. Larkin, “Accelerated Image Reconstruction using Ordered Subsets of Projection Data,” *IEEE Trans Med Imaging*, vol. 13, no. 4, pp. 601–609, 1994.
[41] X. Jin, C. Chan, T. Mulnix, V. Panin, M. E. Casey, C. Liu, and R. E. Carson, “List-mode reconstruction for the Biograph mCT with physics modeling and event-by-

event motion correction.," *Phys. Med. Biol.*, vol. 58, no. 16, pp. 5567–91, Aug. 2013.
[42] S. Mustafovic, K. Thielemans, D. Hogg, and P. Bloomfield, "Object dependency of resolution and convergence rate in OSEM with filtering," *2001 IEEE Nucl. Sci. Symp. Conf. Rec. (Cat. No.01CH37310)*, p. 1786–90 vol.3|4 vol. l+2518, 2002.
[43] G. Mettivier, V. Tabacchini, M. Conti, and P. Russo, "Signal-to-Noise Gain at Variable Randoms Ratio in TOF PET," *IEEE Trans Nucl Sci.*, vol. 59, no. 5, pp. 1948–1957, 2012.
[44] IPEM, "IPEM Report 100: Mathematical Techniques in Nuclear Medicine," 2011.
[45] S. Cherry, J. Sorenson, and M. Phelps, *Physics in Nuclear Medicine*, 3rd Editio. Saunders, 2003.
[46] C. Chan, S. Meikle, R. Fulton, G.-J. Tian, W. Cai, and D. D. Feng, "A non-local post-filtering algorithm for PET incorporating anatomical knowledge," *2009 IEEE Nucl. Sci. Symp. Conf. Rec.*, pp. 2728–2732, Oct. 2009.
[47] GE Healthcare, "Clinical Implementation of VUE Point FX," 2009.
[48] D. Kadrmas, "LOR-OSEM: statistical PET reconstruction from raw line of response histograms," *Phys Med Biol*, vol. 49, no. 20, pp. 4731–4744, 2004.
[49] A. Alessio, C. W. Stearns, S. Tong, S. G. Ross, S. Kohlmyer, A. Ganin, and P. Kinahan, "Application and Evaluation of a Measured Spatially Variant System Model for PET Image Reconstruction," *IEEE Trans Med Imaging*, vol. 29, no. 3, pp. 938–949, 2010.
[50] P. E. Valk, D. Delbeke, D. L. Bailey, D. W. Townsend, and M. N. Maisey, *Positron Emission Tomography: Clinical Practise*, First Edit. 2003.
[51] S. D. Wollenweber, "Parameterization of a Model-Based 3-D PET Scatter Correction," *IEEE Trans Nucl Sci.*, vol. 49, no. 3, pp. 722–727, 2002.
[52] P. F. Sharp, H. G. Gemmell, and A. D. Murray, *Practical Nuclear Medicine*, 3rd Editio. 2005.
[53] M. A. Lodge, A. Rahmim, and R. L. Wahl, "Simultaneous measurement of noise and spatial resolution in PET phantom images.," *Phys. Med. Biol.*, vol. 55, no. 4, pp. 1069–81, Feb. 2010.
[54] C. W. Stearns, D. L. Mcdaniel, S. G. Kohlmyer, P. R. Arul, B. P. Geiser, and V. Shanmugam, "Random Coincidence Estimation from Single Event Rates on the Discovery ST PET / CT Scanner," *IEEE Nucl Sci Symp Conf Rec.*, pp. 3067–3069, 2004.
[55] D. Brasse, P. E. Kinahan, C. Lartizien, C. Comtat, M. Casey, and C. Michel, "Correction methods for random coincidences in fully 3D whole-body PET: impact on data and image quality.," *J. Nucl. Med.*, vol. 46, no. 5, pp. 859–67, May 2005.
[56] C. Bai, L. Shao, A. J. Da Silva, and Z. Zhao, "A Generalized Model for the Conversion From CT Numbers to Linear Attenuation Coefficients," vol. 50, no. 5, pp. 1510–1515, 2003.
[57] A. H. R. Lonn, "Evaluation of method to minimize the effect of X-ray contrast in PET-CT attenuation correction," *IEEE Nucl. Sci. Symp. Conf. Rec.*, vol. 3, no. 6, pp. 2220–2221, 2003.
[58] M. Abella, S. Member, D. Mankof, J. J. Vaquero, S. Member, M. Desco, and P. E. Kinahan, "Accuracy of CT-Based Attenuation Correction in Bone Imaging with PET / CT," *IEEE Nucl. Sci. Symp. Conf. Rec.*, pp. 4485–4488, 2007.
[59] D. Bailey, D. Townsend, P. Valk, and M. Maisey, *Positron Emission Tomography (PET)*. Springer, 2005.
[60] P. Zanzonico, "Routine quality control of clinical nuclear medicine instrumentation: a brief review.," *J. Nucl. Med.*, vol. 49, no. 7, pp. 1114–31, Jul. 2008.
[61] G. B. Saha, *Basics of PET Imaging*. New York, NY: Springer New York, 2010.
[62] R. Boellaard, W. J. G. Oyen, C. J. Hoekstra, O. S. Hoekstra, E. P. Visser, A. T. Willemsen, B. Arends, F. J. Verzijlbergen, J. Zijlstra, A. M. Paans, E. F. I. Comans, and J. Pruim, "The Netherlands protocol for standardisation and quantification of FDG whole body PET studies in multi-centre trials," *Eur. J. Nucl. Med. Mol. Imaging*, vol. 35, no. 12, pp. 2320–2333, 2008.
[63] M. Diksic and R. C. Reba, *Radiopharmaceuticals and Brain Pathophysiology Studied with PET and SPECT*. CRC Press, 1990.

[64] P. Christian and N. Swanston, *SNMTS PET Study Guide*. 2010.
[65] C. W. Stearns, “NEC and Local Image Noise in PET Imaging,” *IEEE Nucl. Sci. Symp. Conf. Rec.*, vol. 5, pp. 3106–3108, 2004.
[66] M. C. Adams, T. G. Turkington, J. M. Wilson, and T. Z. Wong, “A systematic review of the factors affecting accuracy of SUV measurements.,” *AJR. Am. J. Roentgenol.*, vol. 195, no. 2, pp. 310–20, Aug. 2010.
[67] C. Lasnon, C. Desmonts, E. Quak, R. Gervais, P. Do, C. Dubos-Arvis, and N. Aide, “Harmonizing SUVs in multicentre trials when using different generation PET systems: prospective validation in non-small cell lung cancer patients.,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 40, no. 7, pp. 985–96, Jul. 2013.
[68] M. Lodge, M. Chaudhry, and R. L. Wahl, “Noise considerations for PET quantification using maximum and peak standardized uptake value.,” *J. Nucl. Med.*, vol. 53, no. 7, pp. 1041–7, Jul. 2012.
[69] P. Keim, “An Overview of PET Quality Assurance Procedures: Part 1,” *J NuclMed Technol*, vol. 22, no. 1, pp. 27–34, 1994.
[70] National Electrical Manufacturers Association, “NEMA NU 2-2012 Performance Measurements of Positron Emission Tomographs,” 2013.
[71] J. Nuyts, “Unconstrained image reconstruction with resolution modelling does not have a unique solution,” *EJNMMI Phys.*, vol. 1, no. 98, pp. 1–7, 2014.
[72] J. L. Humm, A. Rosenfeld, and A. Del Guerra, “From PET detectors to PET scanners.,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 30, no. 11, pp. 1574–97, Nov. 2003.
[73] M. N. Lonsdale and T. Beyer, “Dual-modality PET/CT Instrumentation - Today and tomorrow,” *Eur. J. Radiol.*, vol. 73, pp. 452–460, 2010.
[74] T. Beyer, D. W. Townsend, J. Czernin, and L. S. Freudenberg, “The future of hybrid imaging-part 2: PET/CT.,” *Insights Imaging*, vol. 2, no. 3, pp. 225–34, Jun. 2011.
[75] M. M. Khalil, *Basic Sciences of Nuclear Medicine*. Springer, 2011.
[76] A. Ferretti, E. Bellan, M. Gava, S. Chondrogiannis, A. Massaro, O. Nibale, and D. Rubello, “Phantom study of the impact of reconstruction parameters on the detection of mini- and micro-volume lesions with a low-dose PET/CT acquisition protocol.,” *Eur. J. Radiol.*, vol. 81, no. 11, pp. 3363–70, Nov. 2012.
[77] M. Soret, S. L. Bacharach, and I. Buvat, “Partial-volume effect in PET tumor imaging.,” *J. Nucl. Med.*, vol. 48, no. 6, pp. 932–45, Jun. 2007.
[78] K. Erlandsson, I. Buvat, P. H. Pretorius, B. a Thomas, and B. F. Hutton, “A review of partial volume correction techniques for emission tomography and their applications in neurology, cardiology and oncology.,” *Phys. Med. Biol.*, vol. 57, no. 21, pp. R119-59, Nov. 2012.
[79] GE Healthcare, “Lightburst Digital Detector Workflow,” 2017.
[80] M. A. Miller, “Vereos White Paper: Focusing on high performance,” 2016.
[81] S. Ross, “Q.Clear White Paper,” 2014.
[82] M. Conti, “Improving time resolution in time-of-flight PET,” *Nucl. Inst. Methods Phys. Res. A*, vol. 648, pp. S194–S198, 2011.
[83] M. Conti, “Focus on time-of-flight PET: the benefits of improved time resolution.,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 38, no. 6, pp. 1147–57, Jun. 2011.
[84] T. Budinger, “Time-of-Flight Positron Emission Tomography: Status Relative to Conventional PET,” *J Nucl Med*, vol. 24, no. 1, pp. 73–78, 1983.
[85] E. Clementel, P. Mollet, and S. Vandenberghe, “Effect of Local TOF Kernel Miscalibrations on Contrast-Noise in TOF PET,” *IEEE Trans. Nucl. Sci.*, vol. 60, no. 3, pp. 1578–1588, Jun. 2013.
[86] E. Rapisarda, V. Bettinardi, K. Thielemans, and M. C. Gilardi, “Image-based point spread function implementation in a fully 3D OSEM reconstruction algorithm for PET,” *Phys Med Biol*, vol. 4131, no. 55, pp. 4131–4151, 2010.
[87] F. C. Sureau, A. J. Reader, C. Comtat, C. Leroy, M.-J. Ribeiro, I. Buvat, and R. Trebossen, “Impact of Image-Space Resolution Modeling for Studies with the High-Resolution Research Tomograph,” *J. Nucl. Med.*, vol. 49, no. 6, pp. 1000–1008, 2008.
[88] V. Y. Panin, F. Kehren, C. Michel, and M. Casey, “Fully 3-D PET reconstruction with system matrix derived from point source measurements,” *IEEE Trans. Med.*

Imaging, vol. 25, no. 7, pp. 907–921, 2006.
[89] S. Ross and C. Stearns, “SharpIR White Paper.”
[90] S. Tong, A. Alessio, and P. Kinahan, “Noise and signal properties in PSF-based fully 3D PET image reconstruction: an experimental evaluation,” *Phys Med Biol*, vol. 1453, no. 55, pp. 1453–1473, 2010.
[91] A. M. Alessio, P. E. Kinahan, and T. K. Lewellen, “Modeling and incorporation of system response functions in 3-D whole body PET,” *IEEE Trans. Med. Imaging*, vol. 25, no. 7, pp. 828–837, 2006.
[92] S. Tong, A. Alessio, and P. Kinahan, “Evaluation of Noise Properties in PSF-Based PET Image Reconstruction,” *IEEE Nucl Sci Symp Conf Rec.*, vol. 2009, no. 2009, pp. 3042–3047, 2009.
[93] A. Rahmim and J. Tang, “Noise propagation in resolution modeled PET imaging and its impact on detectability.,” *Phys. Med. Biol.*, vol. 58, no. 19, pp. 6945–68, Oct. 2013.
[94] K. Gong, S. R. Cherry, and J. Qi, “On the assessment of spatial resolution of PET systems with iterative image reconstruction,” *Phys. Med. Biol.*, vol. 61, no. 5, pp. N193–N202, 2016.
[95] D. L. Snyder, M. I. Miller, L. J. Thomas, and D. G. Politte, “Noise and Edge Artifacts in Maximum-Likelihood Reconstructions for Emission Tomography,” *IEEE Trans. Med. Imaging*, vol. 6, no. 3, pp. 228–238, 1987.
[96] R. Boellaard, “Optimisation and harmonisation: two sides of the same coin?,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 40, no. 7, pp. 982–4, Jul. 2013.
[97] A. Sacks, P. J. Peller, D. S. Surasi, L. Chatburn, G. Mercier, and R. M. Subramaniam, “Value of PET/CT in the management of liver metastases, part 1,” *Am. J. Roentgenol.*, vol. 197, no. 2, pp. 256–259, 2011.
[98] A. Nasoodi, R. Syed, A. Afaq, and J. Bomanji, “Use of PET/MRI for identification and characterisation of liver lesions,” *Clin. Transl. Imaging*, vol. 2, no. 2, pp. 129–137, 2014.
[99] J. Cleary, H. Gelband, and J. Wagner, “Cancer: Disease Control Priorities,” *Dis. Control Priorities 3rd Ed.*, pp. 1–363, 2015.
[100] M. Arnold, M. S. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, “Global patterns and trends in colorectal cancer incidence and mortality,” *Gut*, vol. 66, no. 4, pp. 683–691, 2017.
[101] B. Wiering, W. V Vogel, T. J. M. Ruers, and W. J. G. Oyen, “Controversies in the management of colorectal liver metastases: role of PET and PET/CT.,” *Dig. Surg.*, vol. 25, no. 6, pp. 413–20, 2008.
[102] P. J. Robinson, “The early detection of liver metastases,” *Cancer Imaging*, vol. 2, no. 2, pp. 1–3, 2002.
[103] L.-H. Xu, “Imaging diagnosis of colorectal liver metastases,” *World J. Gastroenterol.*, vol. 17, no. 42, p. 4654, 2011.
[104] D. V Sahani, S. P. Kalva, A. J. Fischman, R. Kadavigere, M. Blake, P. F. Hahn, and S. Saini, “Detection of Liver Metastases from Adenocarcinoma of the Colon and Pancreas: Comparison of Mangafodipir Trisodium – Enhanced Liver MRI and Whole-Body FDG PET,” *Am. J. Roentgenol.*, vol. 185, pp. 239–246, 2005.
[105] M. Yang, D. R. Martin, N. Karabulut, and M. P. Frick, “Comparison of MR and PET imaging for the evaluation of liver metastases,” *J. Magn. Reson. Imaging*, vol. 17, no. 3, pp. 343–349, 2003.
[106] K. Kinkel, Y. Lu, M. Both, R. S. Warren, and R. F. Thoeni, “Detection of hepatic metastases from cancers of the gastrointestinal tract by using noninvasive imaging methods (US, CT, MR imaging, PET): a meta-analysis.,” *Radiology*, vol. 224, no. 3, pp. 748–756, 2002.
[107] G. Grassetto, A. Fornasiero, G. Bonciarelli, E. Banti, L. Rampin, M. C. Marzola, A. Massaro, F. Galeotti, G. Del Favero, F. Pasini, A. M. Minicozzi, A. Al-Nahhas, C. Cordiano, and D. Rubello, “Additional value of FDG-PET/CT in management of ‘Solitary’ liver metastases: Preliminary results of a prospective multicenter study,” *Mol. Imaging Biol.*, vol. 12, no. 2, pp. 139–144, 2010.
[108] R. A. Abbadi, U. Sadat, A. Jah, R. K. Praseedom, N. V. Jamieson, H. K. Cheow, S. Whitley, H. E. Ford, C. B. Wilson, S. J. F. Harper, and E. L. Huguet, “Improved

long-term survival after resection of colorectal liver metastases following staging with FDG positron emission tomography," *J. Surg. Oncol.*, vol. 110, no. 3, pp. 313–319, 2014.
[109] W. H. Luk, A. W. S. Au-Yeung, and T. K. L. Loke, "Imaging patterns of liver uptakes on PET scan: Pearls and pitfalls," *Nucl. Med. Rev.*, vol. 16, no. 2, pp. 75–81, 2013.
[110] S. Surti, J. Scheuermann, G. El Fakhri, M. E. Daube-Witherspoon, R. Lim, N. Abi-Hatem, E. Moussallem, F. Benard, D. Mankoff, and J. S. Karp, "Impact of time-of-flight PET on whole-body oncologic studies: a human observer lesion detection and localization study.," *J. Nucl. Med.*, vol. 52, no. 5, pp. 712–9, May 2011.
[111] B. W. Jakoby, Y. Bercier, M. Conti, M. E. Casey, and B. Bendriem, "Physical and clinical performance of the mCT time-of-flight PET / CT scanner," *Phys Med Biol*, vol. 56, pp. 2375–2389, 2011.
[112] QIBA, "QIBA Profile. FDG-PET/CT as an Imaging Biomarker Measuring Response to Cancer Therapy," 2013.
[113] QIBA, "The QIBA profile for quantitative FDG-PET/CT oncology imaging v2," 2014.
[114] R. Boellaard, "Standards for PET image acquisition and quantitative data analysis.," *J. Nucl. Med.*, vol. 50 Suppl 1, p. 11S–20S, May 2009.
[115] G. Tomasi, F. Turkheimer, and E. Aboagye, "Importance of quantification for the analysis of PET data in oncology: Review of current methods and trends for the future," *Mol. Imaging Biol.*, vol. 14, no. 2, pp. 131–136, 2012.
[116] S. F. Barrington, N. G. Mikhaeel, L. Kostakoglu, M. Meignan, M. Hutchings, S. P. Mueller, L. H. Schwartz, E. Zucca, R. I. Fisher, J. Trotman, O. S. Hoekstra, R. J. Hicks, M. J. O'Doherty, R. Hustinx, A. Biggi, and B. D. Cheson, "Role of imaging in the staging and response assessment of lymphoma: Consensus of the international conference on malignant lymphomas imaging working group," *J. Clin. Oncol.*, vol. 32, no. 27, pp. 3048–3058, 2014.
[117] G. J. S. Tan, S. U. Berlangieri, S. T. Lee, and A. M. Scott, "FDG PET/CT in the liver: Lesions mimicking malignancies," *Abdom. Imaging*, vol. 39, no. 1, pp. 187–195, 2014.
[118] C. Lin, E. Itti, C. Haioun, Y. Petegnief, A. Luciani, J. Dupuis, G. Paone, J.-N. Talbot, A. Rahmouni, and M. Meignan, "Early 18F-FDG PET for prediction of prognosis in patients with diffuse large B-cell lymphoma: SUV-based assessment versus visual analysis.," *J. Nucl. Med.*, vol. 48, no. 10, pp. 1626–1632, 2007.
[119] P. E. Kinahan and J. W. Fletcher, "Positron Emission Tomography-Computed Tomography Standardised Uptake Values in Clinical Practice and Assessing Response to Therapy," *Semin. Ultrasound, CT MRI*, vol. 31, no. 6, pp. 496–505, 2010.
[120] M. Caglar, C. Yener, and E. Karabulut, "Value of CT, FDG PET-CT and serum tumor markers in staging recurrent colorectal cancer," *Int. J. Comput. Assist. Radiol. Surg.*, pp. 993–1002, 2014.
[121] R. Boellaard, A. T. Willemsen, B. Arends, and E. P. Visser, "EARL procedure for assessing PET/CT system specific patient FDG activity preparations for quantitative FDG PET/CT studies," 2010. [Online]. Available: http://earl.eanm.org/html/img/pool/EARL-procedure-for-optimizing-FDG-activity-for-quantitative-FDG-PET-studies_version_1_1.pdf. [Accessed: 01-Jun-2015].
[122] EANM Research Ltd (EARL), "New EANM FDG PET/CT accreditation specifications for SUV recovery coefficients," 2017.
[123] J. J. Sunderland and P. E. Christian, "Quantitative PET/CT Scanner Performance Characterization Based Upon the Society of Nuclear Medicine and Molecular Imaging Clinical Trials Network Oncology Clinical Simulator Phantom," *J. Nucl. Med.*, vol. 56, no. 1, pp. 145–152, 2014.
[124] J. S. Scheuermann, J. R. Saffer, J. S. Karp, A. M. Levering, and B. A. Siegel, "Qualification of PET Scanners for Use in Multicenter Cancer Clinical Trials: The American College of Radiology Imaging Network Experience," *J. Nucl. Med.*, vol. 50, no. 7, pp. 1187–1193, 2009.
[125] E. Busemann Sokole, A. Płachcínska, A. Britten, M. Lyra Georgosopoulou, W. Tindale, and R. Klett, "Routine quality control recommendations for nuclear medicine instrumentation," *Eur. J. Nucl. Med. Mol. Imaging*, vol. 37, no. 3, pp. 662–

671, 2010.
[126] IPEM, “IPEM Report No. 86: Quality Control of Gamma Camera Systems,” 2003.
[127] Medical Devices Agency Gamma Camera Assessment Team, “Gamma Camera Performance Assessment Protocol,” 1996.
[128] Medical Devices Agency Gamma Camera Assessment Team, “Gamma Camera Performance Assessment Protocol Part 2: Wholebody Imaging Systems,” 2001.
[129] S. Surti, A. Kuhn, M. E. Werner, A. E. Perkins, J. Kolthammer, and J. S. Karp, “Performance of Philips Gemini TF PET/CT scanner with special consideration for its time-of-flight imaging capabilities.,” *J. Nucl. Med.*, vol. 48, no. 3, pp. 471–480, 2007.
[130] J. S. Karp, S. Surti, M. E. Daube-Witherspoon, and G. Muehllehner, “Benefit of time-of-flight in PET: experimental and clinical results.,” *J. Nucl. Med.*, vol. 49, no. 3, pp. 462–70, Mar. 2008.
[131] D. J. Kadrmas, M. E. Casey, M. Conti, B. W. Jakoby, C. Lois, and D. W. Townsend, “Impact of time-of-flight on PET tumor detection.,” *J. Nucl. Med.*, vol. 50, no. 8, pp. 1315–23, Aug. 2009.
[132] G. Akamatsu, K. Mitsumoto, T. Taniguchi, Y. Tsutsui, S. Baba, and M. Sasaki, “Influences of point-spread function and time-of-flight reconstructions on standardized uptake value of lymph node metastases in FDG-PET,” *Eur. J. Radiol.*, vol. 83, no. 1, pp. 226–230, 2014.
[133] J. M. Wilson and T. G. Turkington, “TOF-PET Small-Lesion Image Quality Measured Over a Range of Phantom Sizes,” vol. 60, no. 3, pp. 1589–1595, 2013.
[134] G. Akamatsu, K. Ishikawa, K. Mitsumoto, T. Taniguchi, N. Ohya, S. Baba, K. Abe, and M. Sasaki, “Improvement in PET/CT image quality with a combination of point-spread function and time-of-flight in relation to reconstruction parameters.,” *J Nucl Med*, vol. 53, no. 11, pp. 1716–22, Nov. 2012.
[135] E. Prieto, J. M. Martí-Climent, V. Morán, L. Sancho, B. Barbés, J. Arbizu, and J. A. Richter, “Brain PET imaging optimization with time of flight and point spread function modelling,” *Phys. Medica*, vol. 31, no. 8, pp. 948–955, 2015.
[136] K. E. Kenny and D. R. McGowan, “Effect of the GE SharpIR PET reconstruction on contrast convergence and measured SUV values in the NEMA image quality phantom and patients,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 40, no. 2S, pp. S121–S122, 2013.
[137] F. L. Andersen, T. L. Klausen, A. Loft, T. Beyer, and S. Holm, “Clinical evaluation of PET image reconstruction using a spatial resolution model.,” *Eur. J. Radiol.*, vol. 82, no. 5, pp. 862–9, May 2013.
[138] K. Kawashima, K. Kato, M. Tomabechi, M. Matsuo, K. Otsuka, K. Ishida, R. Nakamura, and S. Ehara, “Clinical evaluation of 18 F-fludeoxyglucose positron emission tomography / CT using point spread function reconstruction for nodal staging of colorectal cancer,” *Br J Radiol*, vol. 89, no. 20150938, 2016.
[139] S. Sheikhbahaei, C. Marcus, R. Wray, A. Rahmim, M. A. Lodge, and R. M. Subramaniam, “Impact of point spread function reconstruction on quantitative 18F-FDG-PET/CT imaging parameters and inter-reader reproducibility in solid tumors,” *Nucl. Med. Commun.*, p. 1, 2015.
[140] V. Y. Panin, F. Kehren, H. Rothfuss, D. Hu, C. Michel, M. E. Casey, S. Member, and A. T. H. Scanner, “PET Reconstruction with System Matrix Derived From Point Source Measurements,” *IEEE Trans Nucl Sci.*, vol. 53, no. 1, pp. 152–159, 2006.
[141] C. Lasnon, R. J. Hicks, J.-M. Beauregard, A. Milner, M. Paciencia, and N. Aide, “Impact of Point Spread Function Reconstruction on Thoracic Lymph Node Staging With 18 F-FDG PET / CT in Non Y Small Cell Lung Cancer,” *Clin. Nucl. Med.*, vol. 37, no. 10, pp. 971–976, 2012.
[142] D. Bellevre, C. Blanc Fournier, O. Switsers, A. E. Dugué, C. Levy, D. Allouache, C. Desmonts, H. Crouet, J.-M. Guilloit, J.-M. Grellard, and N. Aide, “Staging the axilla in breast cancer patients with (18)F-FDG PET: how small are the metastases that we can detect with new generation clinical PET systems?,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 41, no. 6, pp. 1103–12, Jun. 2014.
[143] A. Nagaki, M. Onoguchi, and N. Matsutomo, “Clinical validation of high-resolution image reconstruction algorithms in brain 18F-FDG-PET: effect of incorporating

Gaussian filter, point spread function, and time-of-flight.," *Nucl. Med. Commun.*, vol. 35, no. 12, pp. 1224–1232, 2014.
[144] D. Kidera, K. Kihara, G. Akamatsu, S. Mikasa, T. Taniguchi, Y. Tsutsui, T. Takeshita, A. Maebatake, K. Miwa, and M. Sasaki, "The edge artifact in the point-spread function-based PET reconstruction at different sphere-to-background ratios of radioactivity," *Ann. Nucl. Med.*, vol. 30, no. 2, pp. 97–103, 2016.
[145] T. Taniguchi, G. Akamatsu, Y. Kasahara, K. Mitsumoto, S. Baba, Y. Tsutsui, K. Himuro, S. Mikasa, D. Kidera, and M. Sasaki, "Improvement in PET/CT image quality in overweight patients with PSF and TOF," *Ann. Nucl. Med.*, vol. 29, no. 1, pp. 71–77, 2015.
[146] G. Akamatsu, K. Mitsumoto, K. Ishikawa, T. Taniguchi, N. Ohya, S. Baba, K. Abe, and M. Sasaki, "Benefits of point-spread function and time of flight for PET/CT image quality in relation to the body mass index and injected dose.," *Clin. Nucl. Med.*, vol. 38, no. 6, pp. 407–12, Jun. 2013.
[147] R. Matheoud, M. Lecchi, D. Lizio, C. Scabbio, C. Marcassa, L. Leva, A. Del Sole, C. Rodella, L. Indovina, C. Bracco, M. Brambilla, and O. Zoccarato, "Comparative analysis of iterative reconstruction algorithms with resolution recovery and time of flight modeling for 18F-FDG cardiac PET: A multi-center phantom study," *J. Nucl. Cardiol.*, pp. 1–10, 2016.
[148] J. M. Rogasch, F. Hofheinz, A. Lougovski, C. Furth, J. Ruf, O. S. Großer, K. Mohnike, P. Hass, M. Walke, H. Amthauer, and I. G. Steffen, "The influence of different signal-to-background ratios on spatial resolution and F18-FDG-PET quantification using point spread function and time-of-flight reconstruction," *EJNMMI Phys.*, vol. 1, no. 1, p. 12, Sep. 2014.
[149] G. Kuhnert, R. Boellaard, S. Sterzer, D. Kahraman, M. Scheffler, J. Wolf, M. Dietlein, A. Drzezga, and C. Kobe, "Impact of PET/CT image reconstruction methods and liver uptake normalization strategies on quantitative image analysis," *Eur. J. Nucl. Med. Mol. Imaging*, vol. 43, no. 2, pp. 249–258, 2016.
[150] E. J. Teoh, D. R. McGowan, K. M. Bradley, E. Belcher, E. Black, A. Moore, A. Sykes, and F. V. Gleeson, "18F-FDG PET/CT assessment of histopathologically confirmed mediastinal lymph nodes in non-small cell lung cancer using a penalised likelihood reconstruction," *Eur. Radiol.*, pp. 1–9, 2016.
[151] E. J. Teoh, D. R. McGowan, K. M. Bradley, E. Belcher, E. Black, and F. V. Gleeson, "Novel penalised likelihood reconstruction of PET in the assessment of histologically verified small pulmonary nodules," *Eur. Radiol.*, vol. 26, no. 2, pp. 576–584, 2016.
[152] A. Nakamura, Y. Tanizaki, M. Takeuchi, S. Ito, Y. Sano, and M. Sato, "Impact of Point Spread Function Correction in Standardized Uptake Value Quantitation for Positron Emission Tomography Images : A Study," pp. 542–548, 2014.
[153] A. M. Alessio, A. Rahmin, and C. G. Orton, "Resolution modeling enhances PET imaging," *Med. Phys.*, vol. 40, no. 12, pp. 1–4, 2013.
[154] O. L. Munk, L. P. Tolbod, S. B. Hansen, and T. V. Bogsrud, "Point-spread function reconstructed PET images of sub-centimeter lesions are not quantitative," *EJNMMI Phys.*, vol. 4, no. 1, p. 5, 2017.
[155] E. De Ponti, S. Morzenti, L. Guerra, C. Pasquali, M. Arosio, V. Bettinardi, a. Crespi, M. C. Gilardi, and C. Messa, "Performance measurements for the PET/CT Discovery-600 using NEMA NU 2-2007 standards," *Med. Phys.*, vol. 38, no. 2, p. 968, 2011.
[156] J. A. Kolthammer, K. Su, A. Grover, M. Narayanan, and R. F. Muzic, "Performance evaluation of the Ingenuity TF PET/CT scanner with a focus on high count-rate conditions," *Phys. Med. Biol.*, vol. 59, no. 14, pp. 3843–3859, 2014.
[157] A. Morey, F. Noo, and D. Kadrmas, "Effect of using 2mm voxels on observer performance for PET lesion detection," *J. Nucl. Med.*, vol. 55, no. supplement 1, p. 2093, 2014.
[158] D. Koopman, J. A. van Dalen, M. C. M. Lagerweij, H. Arkies, J. de Boer, A. H. J. Oostdijk, C. H. Slump, and P. L. Jager, "Improving the Detection of Small Lesions Using a State-of-the-Art Time-of-Flight PET/CT System and Small-Voxel Reconstructions.," *J. Nucl. Med. Technol.*, vol. 43, no. 1, p. 21–27 7p, 2015.

[159] C. S. van der Vos, D. Koopman, S. Rijnsdorp, A. J. Arends, R. Boellaard, J. A. van Dalen, M. Lubberink, A. T. M. Willemsen, and E. P. Visser, “Quantification, improvement, and harmonization of small lesion detection with state-of-the-art PET,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 44, pp. 4–16, 2017.
[160] J. Zhang, K. Binzel, C. Wright, V. Nagar, P. Bardos, P. Maniawski, and M. V. Knopp, “Does Reducing Reconstruction Voxel Volume / Increasing Matrix Size Improve Lesion Detectability? A phantom and clinical feasibility demonstration using next generation digital PET/CT,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 42, no. Supplement 1, p. 51, 2015.
[161] J. Yan, J. Schaefferkoetter, M. Conti, and D. Townsend, “A method to assess image quality for Low-dose PET: Analysis of SNR, CNR, bias and image noise,” *Cancer Imaging*, vol. 16, no. 1, pp. 1–12, 2016.
[162] D. Delbeke, R. E. Coleman, M. J. Guiberteau, M. L. Brown, H. D. Royal, B. a Siegel, D. W. Townsend, L. L. Berland, J. A. Parker, K. Hubner, M. G. Stabin, G. Zubal, M. Kachelriess, V. Cronin, and S. Holbrook, “Procedure guideline for tumor imaging with 18F-FDG PET/CT 1.0.,” *J. Nucl. Med.*, vol. 47, no. 5, pp. 885–895, 2006.
[163] E. Quak, N. Hovhannisyan, C. Lasnon, C. Fruchart, J. P. Vilque, D. Musafiri, and N. Aide, “The importance of harmonizing interim positron emission tomography in non-Hodgkin lymphoma: Focus on the Deauville criteria,” *Haematologica*, vol. 99, no. 6, 2014.
[164] J. Yan, J. L. Chu-Shern, H. Y. Loi, L. K. Khor, A. K. Sinha, S. T. Quek, I. W. K. Tham, and D. Townsend, “Impact of Image Reconstruction Settings on Texture Features in 18F-FDG PET.,” *J. Nucl. Med.*, vol. 56, no. 11, pp. 1667–73, 2015.
[165] I. S. Armstrong, J. M. James, H. A. Williams, M. D. Kelly, and J. C. Matthews, “The assessment of time-of-flight on image quality and quantification with reduced administered activity and scan times in 18F-FDG PET,” *Nucl Med Commun*, vol. 36, no. 7, pp. 728–737, 2015.
[166] J. M. M. Rogasch, I. G. Steffen, F. Hofheinz, O. S. Großer, C. Furth, K. Mohnike, P. Hass, M. Walke, I. Apostolova, and H. Amthauer, “The association of tumor-to-background ratios and SUVmax deviations related to point spread function and time-of-flight F18-FDG-PET/CT reconstruction in colorectal liver metastases,” *EJNMMI Res.*, vol. 5, no. 1, pp. 1–9, 2015.
[167] GE Healthcare, “Discovery PET/CT 690 Data Sheet.” 2010.
[168] J. M. Wilson and T. G. Turkington, “TOF-PET Small-Lesion Image Quality Measured Over a Range of Phantom Sizes,” *IEEE Nucl Sci Symp Conf Rec.*, pp. 3710–3714, 2009.
[169] R. M. Manjeshwar, S. G. Ross, M. Iatrou, T. W. Deller, C. W. Stearns, and S. Member, “Fully 3D PET Iterative Reconstruction Using Distance- Driven Projectors and Native Scanner Geometry,” *IEEE Nucl Sci Symp Conf Rec.*, pp. 2804–2807, 2006.
[170] J. Cal-Gonzalez, J. L. Herraiz, S. Espana, M. Desco, J. J. Vaquero, and J. M. Udias, “Positron range effects in high resolution 3D PET imaging,” *2009 IEEE Nucl. Sci. Symp. Conf. Rec.*, pp. 2788–2791, 2009.
[171] O. Mawlawi, D. a Podoloff, S. Kohlmyer, J. J. Williams, C. W. Stearns, R. F. Culp, and H. Macapinlac, “Performance characteristics of a newly developed PET/CT scanner using NEMA standards in 2D and 3D modes.,” *J. Nucl. Med.*, vol. 45, no. 10, pp. 1734–1742, 2004.
[172] RStudio Team, “RStudio: Integrated Development for R,” *RStudio, Inc.*, 2016. [Online]. Available: http://www.rstudio.com/.
[173] S. Ross, C. Stearns, R. Manjeshwar, M. Iatrou, and P. I. Ah, “A method of overlap correction for fully 3D OSEM reconstruction of PET data,” *IEEE Nucl Sci Symp Conf Rec.*, vol. 6, pp. 3497–3500, 2004.
[174] D. Visvikis, D. Griffiths, D. C. Costa, J. Bomanji, and P. J. Ell, “Clinical evaluation of 2D versus 3D whole-body PET image quality using a dedicated BGO PET scanner,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 32, no. 9, pp. 1050–1056, 2005.
[175] D. Tout and P. Julyan, “Optimising Slice Overlap in 3D PET-CT Whole Body Acquisitions,” *Nucl. Med. Commun.*, vol. 28, no. 12, p. 961, 2007.

[176] J. Wieseler, G. Johnson, C. Hunt, and P. Peller, “The Significant Impact of Dedicated Head and Neck PET/CT Imaging on Standardized Uptake Values,” *J. Nucl. Med.*, vol. 53, no. supplement 1, p. 2639, May 2012.
[177] NEMA, “NEMA NU 2 Performance Measurements of Positron Emission Tomographs,” 2001.
[178] A. Berkowitz, S. Basu, S. Srinivas, S. Sankaran, S. Schuster, and A. Alavi, “Determination of whole-body metabolic burden as a quantitative measure of disease activity in lymphoma: a novel approach with fluorodeoxyglucose-PET.,” *Nucl. Med. Commun.*, vol. 29, no. 6, pp. 521–6, 2008.
[179] C. McKeown, G. Gillen, M. F. Dempsey, and C. Findlay, “Influence of slice overlap on positron emission tomography image quality,” *Phys. Med. Biol.*, vol. 61, no. 3, 2016.
[180] J. B. Roedl, R. R. Colen, N. S. Holalkere, A. J. Fischman, N. C. Choi, and M. a. Blake, “Adenocarcinomas of the esophagus: Response to chemoradiotherapy is associated with decrease of metabolic tumor volume as measured on PET-CT. Comparison to histopathologic and clinical response evaluation,” *Radiother. Oncol.*, vol. 89, no. 3, pp. 278–286, 2008.
[181] A. M. Maffione, A. Ferretti, G. Grassetto, E. Bellan, C. Capirci, S. Chondrogiannis, M. Gava, M. C. Marzola, L. Rampin, C. Bondesan, P. M. Colletti, and D. Rubello, “Fifteen different (18)F-FDG PET/CT qualitative and quantitative parameters investigated as pathological response predictors of locally advanced rectal cancer treated by neoadjuvant chemoradiation therapy.,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 40, no. 6, pp. 853–64, Jun. 2013.
[182] American College of Radiology Imaging Network, “NCI Centers of Quantitative Imaging Excellence - Manual of Procedures,” 2013.
[183] IAEA, “Quality Assurance for PET and PET/CT Systems,” 2009.
[184] National Physical Laboratory, “Measurement Good Practice Guide No. 93: Protocol for Establishing and Maintaining the Calibration of Medical Radionuclide Calibrators and their Quality Control,” 2006.
[185] T. Chang, G. Chang, S. Kohlmyer, J. W. Clark, E. Rohren, and O. R. Mawlawi, “Effects of injected dose, BMI and scanner type on NECR and image noise in PET imaging,” *Phys. Med. Biol.*, vol. 56, no. 16, pp. 5275–5285, 2011.
[186] E. H. de Groot, N. Post, R. Boellaard, N. R. Wagenaar, A. T. Willemsen, and J. a van Dalen, “Optimized dose regimen for whole-body FDG-PET imaging.,” *EJNMMI Res.*, vol. 3, no. 1, p. 63, Jan. 2013.
[187] P. Geramifar, M. R. Ay, M. S. Zafarghandi, S. Sarkar, G. Loudos, and A. Rahmim, “Investigation of time-of-flight benefits in an LYSO-based PET / CT scanner : A Monte Carlo study using GATE,” *Nucl. Inst. Methods Phys. Res. A*, vol. 641, pp. 121–127, 2011.
[188] S. D. Wollenweber, K. R. Phillips, and C. W. Stearns, “Calculation of noise-equivalent image quality,” *IEEE Nucl Sci Symp Conf Rec.*, no. 4, pp. 2822–2825, 2004.
[189] M. E. Daube-witherspoon, S. Member, S. Matej, M. E. Werner, S. Surti, and J. S. Karp, “Comparison of List-Mode and DIRECT Approaches for Time-of-Flight PET Reconstruction,” *IEEE Trans Med Imaging*, vol. 31, no. 7, pp. 1461–1471, 2012.
[190] M. E. Daube-witherspoon, S. Member, S. Matej, M. E. Werner, S. Surti, and J. S. Karp, “Impact of Resolution Modeling on Accuracy and Precision of Lesion Contrast Measurements,” *IEEE Nucl Sci Symp Conf Rec.*, pp. 4373–4377, 2011.
[191] K. Wangerin, S. Ahn, S. G. Ross, P. E. Kinahan, and R. M. Manjeshwar, “Improving lesion detectability in PET imaging with a penalized likelihood reconstruction algorithm,” *Med. Imaging 2015 Image Perception, Obs. Performance, Technol. Assess.*, vol. 9416, no. October, p. 94160W, 2015.
[192] S. Surti, “Update on Time-of-Flight PET Imaging,” *J Nucl Med*, vol. 56, no. 1, pp. 98–106, 2015.
[193] M. E. Daube-Witherspoon, J. S. Karp, M. E. Casey, F. P. DiFilippo, H. Hines, G. Muehllehner, V. Simcic, C. W. Stearns, L.-E. Adam, S. Kohlmyer, and V. Sossi, “PET performance measurements using the NEMA NU 2-2001 standard.,” *J. Nucl. Med.*, vol. 43, no. 10, pp. 1398–409, Oct. 2002.

[194] G. Delso, S. Fürst, B. Jakoby, R. Ladebeck, C. Ganter, S. G. Nekolla, M. Schwaiger, and S. I. Ziegler, “Performance measurements of the Siemens mMR integrated whole-body PET/MR scanner.,” *J. Nucl. Med.*, vol. 52, no. 12, pp. 1914–22, Dec. 2011.
[195] H. W. de Jong, F. H. van Velden, R. W. Kloet, F. L. Buijs, R. Boellaard, and A. A. Lammertsma, “Performance evaluation of the ECAT HRRT: an LSO-LYSO double layer high resolution, high sensitivity scanner,” *Phys Med Biol*, vol. 52, no. 5, pp. 1505–1526, 2007.
[196] J.-S. S. Liow and S. C. Strother, “The convergence of object dependent resolution in maximum likelihood based tomographic image reconstruction.,” *Phys. Med. Biol.*, vol. 38, pp. 55–70, 1993.
[197] T. Murata, K. Miwa, N. Miyaji, K. Wagatsuma, T. Hasegawa, K. Oda, T. Umeda, T. Iimori, Y. Masuda, T. Terauchi, and M. Koizumi, “Evaluation of spatial dependence of point spread function-based PET reconstruction using a traceable point-like 22Na source,” *EJNMMI Phys.*, vol. 3, no. 1, p. 26, 2016.
[198] E. Prieto, J. M. Martí-Climent, J. Arbizu, P. Garrastachu, I. Domínguez, G. Quincoces, M. J. García-Velloso, P. Lecumberri, M. Gómez-Fernández, and J. A. Richter, “Evaluation of spatial resolution of a PET scanner through the simulation and experimental measurement of the recovery coefficient,” *Comput. Biol. Med.*, vol. 40, no. 1, pp. 75–80, 2010.
[199] A. Skretting, “A METHOD FOR ON-SITE MEASUREMENTS OF THE EFFECTIVE SPATIAL RESOLUTION IN PET IMAGE VOLUMES RECONSTRUCTED WITH OSEM AND GAUSSIAN POST-FILTERS,” *Radiat. Prot. Dosimetry*, vol. 139, no. 1, pp. 195–198, 2010.
[200] I. Castiglioni, G. Rizzo, M. C. Gilardi, V. Bettinardi, A. Savi, F. Fazio, and A. M. C. S. Model, “Lesion Detectability and Quantification in PET / CT Oncological Studies by Monte Carlo Simulations,” *IEEE Trans Nucl Sci.*, vol. 52, no. 1, pp. 136–142, 2005.
[201] A. Suljic, P. Tomse, L. Jensterle, and D. Skrk, “The impact of reconstruction algorithms and time of flight information on PET/CT image quality,” *Radiol. Oncol.*, vol. 49, no. 3, pp. 227–233, 2015.
[202] J. A. Stamos, W. L. Rogers, N. H. Clinthorne, and K. F. Koral, “Object-dependent performance comparison of two iterative reconstruction algorithms,” *IEEE Trans. Nucl. Sci.*, vol. 35, no. 1, pp. 611–614, 1988.
[203] K. Erlandsson, K. Kacperski, D. van Gramberg, and B. F. Hutton, “Performance evaluation of D-SPECT: a novel SPECT system for nuclear cardiology.,” *Phys. Med. Biol.*, vol. 54, no. 9, pp. 2635–49, 2009.
[204] D. Badger and L. Barnden, “Spatial resolution is dependent on image content for SPECT with iterative reconstruction incorporating distance dependent resolution (DDR) correction,” *Australas. Phys. Eng. Sci. Med.*, vol. 37, no. 3, pp. 551–557, 2014.
[205] C. Brown, C. Findlay, A. Nicol, and G. Gillen, “Practical Methods for Measuring Clinical SPECT Spatial Resolution,” *Nucl Med Commun*, vol. 37, p. 559, 2016.
[206] B. De Man and S. Basu, “Distance-driven projection and backprojection in three dimensions,” *Phys Med Biol*, vol. 2463, no. 49, pp. 2463–2475, 2004.
[207] M. A. Lodge, A. Rahmin, and R. L. Wahl, “A Practical, Automated Quality Assurance Method for Measuring Spatial Resolution in PET,” *J Nucl Med*, vol. 50, no. 8, pp. 1307–1314, 2009.
[208] O. V. Olesen, M. Sibomana, S. H. Keller, F. Andersen, J. Jensen, S. Holm, C. Svarer, and L. Højgaard, “Spatial resolution of the HRRT PET scanner using 3D-OSEM PSF reconstruction,” *IEEE Nucl. Sci. Symp. Conf. Rec.*, pp. 3789–3790, 2009.
[209] X. Qin, J. Xu, and Y. Zhong, *Multidisciplinary Management of Liver Metastases in Colorectal Cancer: Early Diagnosis and Treatment*. Springer, 2017.
[210] K. Murakami, “FDG-PET for hepatobiliary and pancreatic cancer: Advances and current limitations.,” *World J. Clin. Oncol.*, vol. 2, no. 5, pp. 229–36, 2011.
[211] A. Mehranian, F. Kotasidis, and H. Zaidi, “Accelerated time-of-flight (TOF) PET image reconstruction using TOF bin subsetization and TOF weighting matrix pre-

computation," *Phys. Med. Biol.*, vol. 61, no. 3, pp. 1309–1331, 2016.
[212] I. S. Armstrong, M. D. Kelly, H. A. Williams, and J. C. Matthews, "Impact of point spread function modelling and time of flight on FDG uptake measurements in lung lesions using alternative filtering strategies," *EJNMMI Phys.*, vol. 1, no. 1, pp. 1–18, 2014.
[213] A. M. Morey and D. J. Kadrmas, "Effect of varying number of OSEM subsets on PET lesion detectability.," *J. Nucl. Med. Technol.*, vol. 41, no. 4, pp. 268–73, Dec. 2013.
[214] G. El Fakhri, S. Surti, C. M. Trott, J. Scheuermann, and J. S. Karp, "Improvement in lesion detection with whole-body oncologic time-of-flight PET.," *J. Nucl. Med.*, vol. 52, no. 3, pp. 347–53, Mar. 2011.
[215] H. Bal, L. Guerin, M. E. Casey, M. Conti, L. Eriksson, C. Michel, S. Fanti, C. Pettinato, S. Adler, and P. Choyke, "Improving PET spatial resolution and detectability for prostate cancer imaging," *Phys. Med. Biol.*, vol. 59, no. 15, pp. 4411–4426, 2014.
[216] X. Zhang, Q. Peng, J. Zhou, J. S. Huber, W. W. Moses, and J. Qi, "Lesion detection and quantification performance of the Tachyon-I time-of-flight PET scanner: phantom and human studies," 2018.
[217] D. J. Kadrmas, M. B. Oktay, M. E. Casey, S. Member, and J. J. Hamill, "Effect of Scan Time on Oncologic Lesion Detection in Whole Body PET," *IEEE Trans Nucl Sci.*, vol. 59, no. 5, pp. 1940–1947, 2012.
[218] D. J. Kadrmas, "The Utah PET lesion detection database," *2009 IEEE Nucl. Sci. Symp. Conf. Rec.*, pp. 3693–3699, Oct. 2009.
[219] S. Kulkarni, P. Khurd, I. Hsiao, L. Zhou, and G. Gindi, "A Channelized Hotelling Observer Study of Lesion Detection in SPECT MAP Reconstruction using Anatomical Priors," *Phys Med Biol*, vol. 52, no. 12, pp. 3601–3617, 2007.
[220] D. J. Kadrmas, T. J. Bradshaw, C. Seegmiller, M. E. Casey, and J. J. Hamill, "Scan time reduction with advanced PET reconstruction: Preserving lesion detection performance," *IEEE Nucl. Sci. Symp. Conf. Rec.*, pp. 2381–2387, 2010.
[221] T. Chang, G. Chang, J. W. Clark, R. H. Diab, E. Rohren, and O. R. Mawlawi, "Reliability of predicting image signal-to-noise ratio using noise equivalent count rate in PET imaging," *Med. Phys.*, vol. 39, no. 10, p. 5891, 2012.
[222] P. E. Kinahan and J. S. Karp, "Figures of merit for comparing reconstruction algorithms with a volume-imaging PET scanner," *Phys. Med. Biol.*, vol. 39, no. 3, pp. 631–642, 1994.
[223] S. Adler, J. Seidel, P. Choyke, M. V. Knopp, K. Binzel, J. Zhang, C. Barker, S. Conant, and R. Maass-Moreno, "Minimum lesion detectability as a measure of PET system performance," *EJNMMI Phys.*, vol. 4, no. 1, 2017.
[224] A. Rahmim, M. A. Lodge, J. Tang, S. Lashkari, and M. R. Ay, "Analytic system matrix resolution modeling in PET: An application to Rb-82 cardiac imaging," *Phys. Med. Biol.*, vol. 53, no. 21, pp. 5947–5965, 2008.
[225] M. Iatrou, S. G. Ross, R. M. Manjeshwar, and C. W. Stearns, "A fully 3D iterative image reconstruction algorithm incorporating data corrections," *IEEE*, vol. 00, no. C, pp. 2493–2497, 2004.
[226] G. Wang and J. Qi, "PET Image Reconstruction Using Kernel Method," *Med. Imaging, IEEE Trans.*, vol. 34, no. 1, pp. 61–71, 2014.
[227] S. Stute and C. Comtat, "Practical considerations for image-based PSF and blobs reconstruction in PET.," *Phys. Med. Biol.*, vol. 58, no. 11, pp. 3849–70, Jun. 2013.
[228] NEMA, "NEMA IEC Body Phantom Set ™." 2001.
[229] H. Fukukita, K. Suzuki, K. Matsumoto, T. Terauchi, H. Daisaki, Y. Ikari, N. Shimada, and M. Senda, "Japanese guideline for the oncology FDG-PET/CT data acquisition protocol: Synopsis of Version 2.0," *Ann. Nucl. Med.*, vol. 28, no. 7, pp. 693–705, 2014.
[230] Office for National Statistics, "'Average' Briton highlighted on UN World Statistics Day," *Econ. Labour Mark. Rev.*, vol. 4, no. 11, p. 6, 2010.
[231] P. M. DeLuca, A. Wambersie, and G. F. Whitmore, "Report 79: Receiver Operating Characteristic Analysis in Medical Imaging," *J. ICRU*, vol. 8, no. 1, pp. 37–40, 2008.
[232] S. D. Wollenweber, A. M. Alessio, and P. E. Kinahan, "A phantom design for

assessment of detectability in PET imaging," *Med. Phys.*, vol. 43, no. 9, pp. 5051–5062, 2016.

[233] J. Sjovall, "PET in the evaluation of head and neck cancer treatment management of the neck," 2015.

[234] C.-B. Kong, B. H. Byun, I. Lim, C. W. Choi, S. M. Lim, W. S. Song, W. H. Cho, D.-G. Jeon, J.-S. Koh, J. Y. Yoo, and S.-Y. Lee, "18F-FDG PET SUVmax as an indicator of histopathologic response after neoadjuvant chemotherapy in extremity osteosarcoma.," *Eur. J. Nucl. Med. Mol. Imaging*, vol. 40, no. 5, pp. 728–36, May 2013.

[235] J. Mertens, A. Dobbeleir, H. Ham, Y. D'Asseler, I. Goethals, and C. Van de Wiele, "Standardized added metabolic activity (SAM): a partial volume independent marker of total lesion glycolysis in liver metastases.," *Eur. J. Nucl. Med. Mol. Imaging*, vol. 39, no. 9, pp. 1441–8, Sep. 2012.

[236] M. Lapela, a Eigtved, S. Jyrkkiö, R. Grénman, T. Kurki, P. Lindholm, J. Nuutinen, E. Sutinen, O. Solin, I. Bjornskov, P. Bretlau, L. Friberg, S. Holm, M. Jensen, H. Sand Hansen, and H. Minn, "Experience in qualitative and quantitative FDG PET in follow-up of patients with suspected recurrence from head and neck cancer.," *Eur. J. Cancer*, vol. 36, no. 7, pp. 858–67, May 2000.

[237] D. P. Chakraborty, "Recent advances in observer performance methodology: Jackknife free-response ROC (JAFROC)," *Radiat. Prot. Dosimetry*, vol. 114, no. 1–3, pp. 26–31, 2005.

[238] D. P. Chakraborty, "The FROC, AFROC and DROC Variants of the ROC Analysis," in *Handbook of Medical Imaging*, vol. 1, 2000, pp. 771–796.

[239] F. Zanca, F. Claus, C. Van Ongeval, V. Provoost, J. Jacobs, G. Marchal, H. Bosmans, and D. P. Chakraborty, "Correlation of Froc and Roc Figures of Merit : Results From Independently Conducted Froc / Roc Studies in Mammography," *Med. Phys.*, vol. 39, no. 10, pp. 5917–5929, 2012.

[240] D. P. Chakraborty, *Observer Performance Methods for Diagnostic Imaging: Foundations, Modeling, and Applications with R-Based Examples*, 1st Editio. CRC Press, 2018.

[241] C. Lartizien, P. E. Kinahan, and C. Comtat, "A Lesion Detection Observer Study Comparing 2-Dimensional Versus Fully 3-Dimensional Whole-Body PET Imaging Protocols," *J. Nucl. Med.*, vol. 45, no. 4, pp. 714–723, 2004.

[242] C. Lartizien, P. E. Kinahan, R. Swensson, C. Comtat, M. Lin, V. Villemagne, and R. Trébossen, "Evaluating image reconstruction methods for tumor detection in 3-dimensional whole-body PET oncology imaging.," *J. Nucl. Med.*, vol. 44, no. 2, pp. 276–90, Feb. 2003.

[243] R. Hustinx, R. J. Dolin, F. Benard, A. Bhatnagar, D. Chakraborty, R. J. Smith, S. Jang, and A. Alavi, "Impact of attenuation correction on the accuracy of FDG-PET in patients with abdominal tumors: A free-response ROC analysis," *Eur. J. Nucl. Med.*, vol. 27, no. 9, pp. 1365–1371, 2000.

[244] A. Teramoto, H. Fujita, K. Takahashi, O. Yamamuro, T. Tamaki, M. Nishio, and T. Kobayashi, "Hybrid method for the detection of pulmonary nodules using positron emission tomography/computed tomography: A preliminary study," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 9, no. 1, pp. 59–69, 2014.

[245] T. H. Farquhar, J. Llacer, C. K. Hoh, J. Czernin, S. S. Gambhir, M. a Seltzer, D. H. Silverman, J. Qi, C. Hsu, and E. J. Hoffman, "ROC and localization ROC analyses of lesion detection in whole-body FDG PET: effects of acquisition mode, attenuation correction and reconstruction algorithm.," *J. Nucl. Med.*, vol. 40, no. 12, pp. 2043–52, 1999.

[246] L. R. MacDonald, D. S. Hippe, L. C. Bender, E. W. Cotter, P. R. Voria, P. S. Hallam, C. L. Wang, D. R. Haseley, M. M. Kelly, J. R. Parikh, J. D. Beatty, and J. V. Rogers, "Positron Emission Mammography Image Interpretation for Reduced Image Count Levels," *J. Nucl. Med.*, vol. 57, no. 3, pp. 348–354, 2016.

[247] S. N. Chatziioannou, A. T. Georgakopoulos, N. K. Pianou, G. T. Kafiri, S. N. Pavlou, and M. Kallergi, "Recurrent Thyroid Cancer Diagnosis. ROC Study of the Effect of a High-Resolution Head and Neck18F-FDG PET/CT Scan.," *Acad. Radiol.*, vol. 21, no. 1, pp. 58–63, 2014.

[248] A. I. Valderrama-Treviño, B. Barrera-Mera, J. C. Ceballos-Villalva, and E. E. Montalvo-Javé, “Hepatic Metastasis from Colorectal Cancer.,” *Euroasian J. hepato-gastroenterology*, vol. 7, no. 2, pp. 166–175, 2017.
[249] J. R. Landis and G. G. Koch, “The Measurement of Observer Agreement for Categorical Data,” *Biometrics*, vol. 33, no. 1, p. 159, 1977.